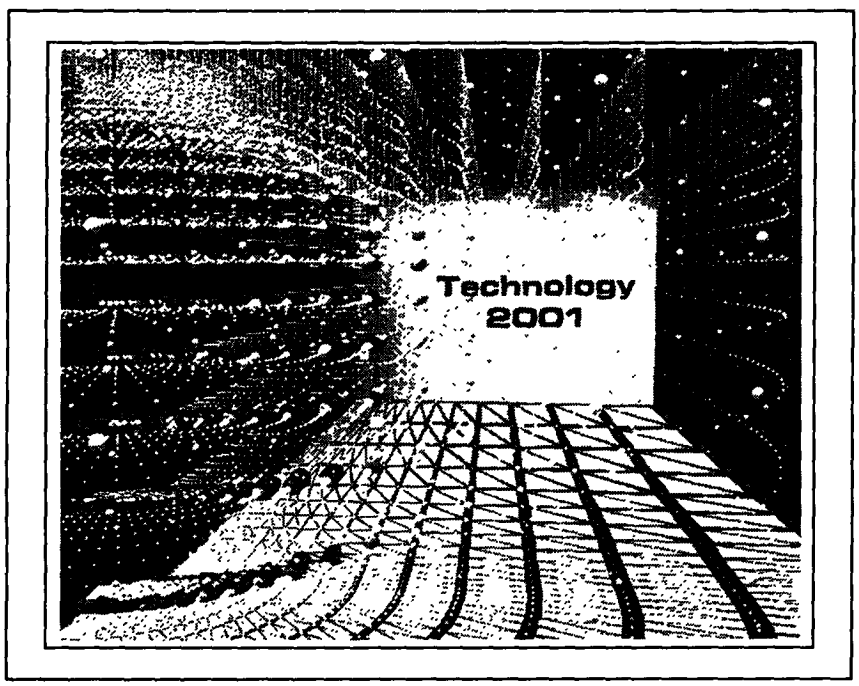
TECHNOLOGY 2001

The Second National Technology Transfer Conference and Exposition

December 3-5, 1991
San Jose Convention Center
San Jose, CA



Conference Proceedings

Sponsored by NASA, the Technology Utilization Foundation, and *NASA Tech Briefs Magazine*

(NASA-CP-3136-Vol-2) TECHNOLOGY 2001: THE SECOND NATIONAL TECHNOLOGY TRANSFER CONFERENCE AND EXPOSITION, VOLUME 2 (NASA) 518 p CSCL 05B

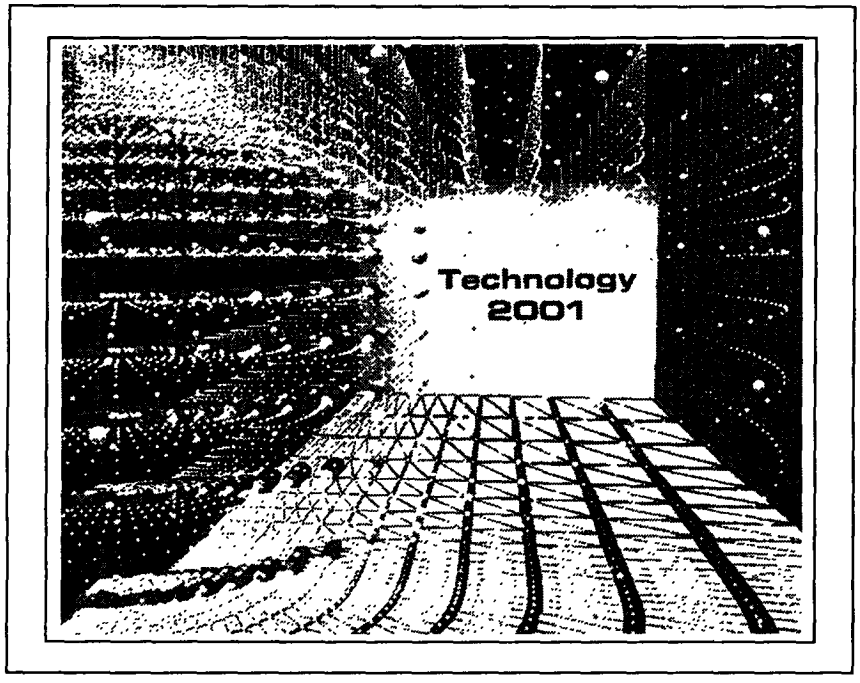
N92-22676
--THRU--
N92-22735
Unclas
H1/99 0073665

NASA Conference Publication 3136, Vol. 2

TECHNOLOGY 2001

The Second National Technology Transfer Conference and Exposition

December 3-5, 1991
San Jose Convention Center
San Jose, CA



Conference Proceedings

Sponsored by NASA, the Technology Utilization Foundation, and *NASA Tech Briefs Magazine*

TECHNOLOGY 2001 - SYMPOSIA PROCEEDINGS

Presented December 3-5, 1991
San Jose, California

TECHNOLOGY 2001 was the second national technology transfer conference and exposition. Held at the San Jose Convention Center December 3-5, 1991, TECHNOLOGY 2001 built upon the foundation laid by last year's initial conference in Washington, D.C., the mission being to transfer advanced technologies developed by the Federal government, its contractors, and other high-tech organizations to U.S. industries for their use in developing new or improved products and processes.

TECHNOLOGY 2001 was sponsored by the National Aeronautics and Space Administration (NASA), *NASA Tech Briefs* magazine, and the Technology Utilization Foundation, with the participation of the following Federal agencies:

Department of Agriculture	Department of Commerce
Department of Defense	Department of Energy
Department of Health and Human Services	Department of the Interior
Department of Transportation	Department of Veteran Affairs
Environmental Protection Agency	National Science Foundation

In addition to an exhibit showcasing the products and technologies available for sale or license from over 200 exhibitors, this year's conference featured 30 concurrent technical sessions in which 120 papers were presented, agency workshops, industry briefings, and the annual Intelligent Processing Equipment (IPE) Conference held concurrently with TECHNOLOGY 2001.

We are pleased to provide the proceedings from the 30 concurrent sessions. This year's program featured symposia on Advanced Manufacturing, Artificial Intelligence, Biotechnology, Computer Graphics and Simulation, Communications, Data and Information Management, Electronics, Electro-Optics, Environmental Technology, Life Sciences, Materials Science, Medical Advances, Robotics, Software Engineering, and Test and Measurement.

The proceedings have been published in two volumes. Volume One contains the first 60 papers presented (in order), while Volume Two contains the last 60 papers presented (again, in order). Proceedings from the IPE Conference Symposia are published under separate cover.

This is Volume Two. Again, the papers appear in the order in which they were presented at TECHNOLOGY 2001. For information regarding additional copies, please contact:

THE TECHNOLOGY UTILIZATION FOUNDATION
41 East 42nd Street, #921
New York, NY 10017

PRECEDING PAGE BLANK NOT FILMED

TECHNOLOGY 2001 SYMPOSIA PROCEEDINGS - VOLUME TWO

Table of Contents

Wednesday December 4th

Materials Sciences:

Advanced Composite Materials and Processes 3
RTM: Cost-Effective Processing of Composite Structures 12
A Low-Cost Method of Testing Compression-After-Impact Strength of Composite Laminates . 22
Resonant Acoustic Determination of Complex Elastic Moduli 32

Robotics:

A Unique Cable Robot for Space and Earth 43
A Lightweight, High-Strength Dexterous Manipulator Arm 51
Real-Time Interactive Simulator System for Telepresence 60
A Hazard Control System for Robot Manipulations 71

Test and Measurement:

Knowledge-Based Autonomous Test Engineer (KATE) 83
Advanced Computed Tomography Inspection System (ACTIS) 93
High-Resolution Ultrasonic Spectroscopy System for Nondestructive Evaluation 101
Force Limited Vibration Testing 105

Thursday December 5th

Advanced Manufacturing:

Development of a Rotary Joint Fluid Coupling for Space Station Freedom 119
Spline Screw Comprehensive Fastening Strategy 129
Commercial Application of an Innovative Nut Design 136
Inflatable Traversing Probe Seal 140

Artificial Intelligence:

CLIPS: An Expert System Building Tool 149
Fuzzy Logic Applications to Expert Systems and Control 159
Neural Network Technologies 169
From Biological Neural Networks to Thinking Machines 174

Biotechnology:

The Microassay on a Card -- A Rugged, Portable Immunoassay 183
Detection of Small Molecules with a Flow Immunosensor 191
Nucleic Acid Probes in Diagnostic Medicine 197
The Rotating Spectrometer: New Biotechnology for Cell Separation 202

Electronics:

Method for Producing High-Quality Oxide Films on Surface 215
Advanced Silicon on Insulator Technology 225
High-Temperature Superconducting Stripline Filter 232
An Adjustable rf Tuning Element for Microwave, Millimeter Wave, and Submillimeter Wave Circuits 239

PRECEDING PAGE BLANK NOT FILMED

Materials Science:

Passive Chlorophyll Detector 249
Commercial Application of Thermal Protection System Technology 250
Oxynitride Glass Fibers 258
Commercial Applications of Advanced Photovoltaic Technologies 265

Software Engineering:

Software Reengineering 277
COSTMODL: An Automated Software Development Cost Estimation Tool 287
Increasing Productivity Through Total Reuse Management 294
How Hypermedia Can Increase the Productivity of Software Development Teams 301

Advanced Manufacturing:

Intelligent Robotic System with Dual-Arm Dexterous Coordination and Real-Time Vision 311
Neural Network Software for Distortion-Invariant Object Recognition 325
Constraint-Based Scheduling 335
COMPASS: A General Purpose Computer-Aided Scheduling Tool 352

Data and Information Management:

ELAS: A Powerful General-Purpose Image Processing Software 361
TAE Plus: A NASA Tool for Building and Managing Graphical User Interfaces 366
Instrumentation, Performance Visualization, and Debugging Tools for Multiprocessors 377
Operations Automation Using Temporal Dependency Networks 386

Electronics:

Thermoacoustic Refrigeration 397
Ambient Temperature Recorder 407
Fiber-Optic Push-Pull Sensor Systems 415
Commercial Capaciflector 423

Environmental Technology:

Water Quality Monitor 437
Remote Semi-Continuous Flowrate Logging Seepage Meter 445
Calcification Prevention Tablets 452
Automated Carbon Dioxide Cleaning System 459

Materials Science:

Applications of Biologically-Derived Microstructures 469
Structural Modification of Polysaccharides: A Biochemical/Genetic Approach 480
Cryogenic Focusing, Ohmically Heated On-Column Trap 486
Study of the Effect of Hydrocarbon Contamination on PFTE Exposed to Atomic Oxygen . . . 495

Medical Advances:

Applications of the Strategic Defense Initiative's Compact Accelerator Technology 503
Acoustically-Based Fetal Heart Rate Monitor 512
Surgical Force Detection Probe 518
Dynamic Inter-Limb Resistance Exercise Device 533

MATERIALS SCIENCE

(Session C4/Room A2)

Wednesday December 4, 1991

- **Advanced Composite Materials and Processes**
- **RTM: Cost-Effective Processing of Composite Structures**
- **A Low-Cost Method of Testing Compression-After-Impact Strength of Composite Laminates**
- **Resonant Acoustic Determination of Complex Elastic Moduli**

N92-22677

ADVANCED COMPOSITE MATERIALS AND PROCESSES

Robert M. Baucom
Group Leader, Composites Processing
NASA-Langley Research Center
Hampton, VA 23665

ABSTRACT

Composites are generally defined as two or more individual materials, which, when combined into a single material system results in improved physical and/or mechanical properties. The freedom of choice of the starting components for composites allows the generation of materials that can be specifically tailored to meet a variety of applications. Advanced composites are described as a combination of high strength fibers and high performance polymer matrix materials. These advanced materials are required to permit future aircraft and spacecraft to perform in extended environments. Advanced composite precursor materials, processes for conversion of these materials to structures, and selected applications for composites are reviewed.

INTRODUCTION

The idea of combining two or more materials to form a single composite material system which has desirable physical and mechanical properties has been around for centuries. Tools, weapons, houses, ornaments etc. were constructed of combinations of naturally occurring materials. From these primitive origins composites have evolved into very sophisticated materials which utilize fiber reinforced polymer matrices which are routinely used to fabricate high performance aircraft and spacecraft, automobiles, sporting goods, medical appliances, construction products, and a host of other items. The widespread use of advanced composites is attributable to the ability to design material systems which possess unique properties that satisfy the requirement to perform in specific environments. Desirable characteristics of advanced composites include high strength, lightweight, corrosion resistance, long fatigue life, damage tolerance, and manufacturing economy.

Advanced composites have two generally recognized precursors: high strength fibers and high performance polymer matrix binding materials. The most common high strength fibers for advanced composites are kevlar, graphite and fiberglass whereas the polymer matrices are normally polyester, epoxy or polyimides. The specific requirements of the application for advanced composites dictates the selection of the fiber/matrix combination. After the fiber and matrix are selected they are combined into a single system referred to as prepreg. The prepreg is then assembled into the appropriate shape and thickness and subjected to processing which converts the prepreg to a finished part. This review will include currently available fibers and matrices, processing methodology and selected applications for composites.

ADVANCED COMPOSITES

Fiber Materials

Continuous or long staple high strength, low density, small diameter filaments are generally considered the predominant reinforcement for advanced composite materials. Graphite, kevlar, and fiberglass are the most widely utilized fiber systems for these materials. Selected properties for these fibers are shown in Figure 1.

	GRAPHITE	KEVLAR	FIBERGLASS	STEEL	ALUMINUM
TENSILE STRENGTH, Ksi	400-800	350-450	500-600	200-300	75-150
TENSILE MODULUS, Ksi	20-100	18	12	30	10
DENSITY, gm/cc	1.8	1.5	2.5	9.4	3.4
DIAMETER	0.0003	0.0005	0.0004	--	--

FIGURE 1. FIBER PROPERTIES

PRECEDING PAGE BLANK NOT FILMED

The tensile strength and tensile modulus of graphite, kevlar and fiberglass fibers vary depending on the chemical composition of the starting materials, processing conditions, and heat treatment during and after fiber formation. Each of these filament types are generally stronger and lighter than conventional metals such as carbon steel and aluminum. The modulus or stiffness of the same fibers is about half that of steels and about double aluminum. Along with the desirable tensile properties of graphite, kevlar, and fiberglass the weight per unit volume for these fiber systems is substantially lower than steel and aluminum. This combination of high strength and low density is very attractive from a structural efficiency point of view.

Matrix Materials

In order to take advantage of the desirable properties of high performance fibers it is necessary to stabilize the individual filaments with matrix materials. Selection of matrix materials is usually made based on the ease of fabrication, cost, and end-use temperature of the structure to be manufactured. Matrix materials for advanced composites can be metallic or non-metallic. In the aerospace industry organic matrix materials are utilized in the fabrication of a wide variety of structural components for commercial and military aircraft and spacecraft. Polyester and epoxy resin matrix materials are utilized for applications that require retention of strength for moderate periods of time at exposure temperatures under 350°F (see Figure 2). Polyimide resins can be used for applications at temperatures up to 550°F. Although, not used in large quantities, selected metals such as aluminum can be used in applications where strength retention at temperatures up to 700°F is a requirement.

RESIN	REASONABLE USE TEMPERATURE, °C (°F)
POLYESTER	121 (250)
EPOXY	177 (350)
POLYIMIDE	288 (550)
ALUMINUM	371 (700)

FIGURE 2. MATRIX MATERIALS

Composite Material Fabrication

Advanced composite materials are generated by combining reinforcing fibers with matrix materials with one of a wide variety of techniques to make composite prepreg. The selection of the method to create composite prepreg is based on many factors including: the ability to place polymers in solution, the viscosity of the polymer, melt-flow characteristics, fiber forming limitations, resin chemoviscosity, economy, and solvent extraction. The most widely utilized procedures for producing composite prepreg are shown in Figure 3.

COMPOSITE FABRICATION METHOD	FEATURES
Solution Coating	Matrix resin dissolved in solvent. Solvent removal can be difficult.
Hot Melt	Widely used for epoxy prepregs. Industry standard prepreg method.
RTM and RIM	Used to infiltrate preforms. Voids and incomplete wet out are concerns.
Intermingled Filaments	Matrix resin in filament form combined with reinforcement fibers.
Powder	Solvent free. Potential for economical, large volume production.
Slurry	Can use non-organic carrier materials.

FIGURE 3. COMPOSITE FABRICATION TECHNIQUES

ADVANCED COMPOSITE PROCESSING PROCEDURES

Many factors are considered when selecting the appropriate assembly and cure procedure for the fabrication of advanced composite structural articles. In order to select the most efficient processing method for a particular resin/fiber combination a basic understanding of the behavior of the resin matrix material under the influence of temperature and pressure is required. In addition, the ease of assembly of the component, economy of the process, surface finish required, and requirement for secondary operations must be understood. Selected composite processing techniques are detailed hereafter.

Autoclave

Autoclaves are widely used to fabricate complex advanced composite structures. Autoclaves are pressure vessels which range in size from very small to over 100 feet in length. Pressures up to 1,000 psi and temperatures up to 1,200°F can be generated in properly designed autoclaves, but for the fabrication of typical aircraft and spacecraft structures pressure and temperature in the range of 100 psi and 350°F, respectively, are routinely employed. After the composite material is assembled for fabrication of a particular structure, the lay-up is placed in a vacuum bag and inserted in the autoclave (see Figure 4). The vacuum line inside the autoclave is connected to the bag containing the composite part. Appropriate heat and pressure are applied to the part to effect complete consolidation and cure. The part is cooled down and removed from the bag. Autoclaves can be utilized to process a wide range of material systems including polyesters, epoxies, polyimides, bismaleimides, and others. In the aerospace industry structural parts ranging in size from a few inches to over 50 feet long are routinely processed in autoclaves.

Heated Platen Presses

Hydraulically actuated heated platen presses are routinely used to fabricate flat composite laminates (see Figure 5). Very high pressures (>5,000 psi) and high temperatures can be applied with this equipment to cure composite laminates. Heated presses are frequently used to fabricate mechanical test samples including tensile, flexure and short beam laminates. Computer controlled ramp temperature and pressure functions allow precise cure cycles to be maintained for proper laminate fabrication. Hydraulic presses can be equipped with vacuum chambers, cooling manifolds, and displacement monitors to assist in the processing of various composite parts.

Trapped Rubber Processing

Silicone rubber expands rapidly when heat is applied. This characteristic combined with its' excellent high temperature properties permit the fabrication of complex composite parts with economical tooling systems (see Figure 6). Silicone rubber material is first mixed and poured into a cavity which contains the master pattern of the part to be fabricated. The rubber is allowed to chemically cure and the master for the part is removed. The composite material is then laid up with the proper fiber orientation and overall thickness. This assembly is then inserted into the cavity in the cured silicone rubber tool. The cast rubber block with the composite lay-up inside is placed in a pressure containment box and the entire assembly is heated up in the oven. As the heat is introduced into the rubber it expands, which in turn applies pressure to the composite part. The temperature of the oven is increased until the cure temperature for the composite material is reached. If the tool is properly designed the pressure can be accurately controlled at the cure temperature. After the appropriate time at temperature, the entire assembly is cooled down which causes the rubber to shrink away from the fully cured part. The finished composite part is then removed. An example of a composite airfoil fabricated in this manner is shown in Figure 7.

Resin Transfer Molding

Resin Transfer Molding (RTM) is a process in which low viscosity resin is pumped into dry fiber preforms contained in a tooling cavity. The individual layers of the dry fiber material is usually in cloth form which is cut and assembled into the desired shape and thickness. This preform is inserted into a tooling cavity which has the final shape of the structural part. Resin is pumped into the preform with a combination of vacuum and pressure until the preform is fully saturated. This assembly is then heated up to cure the matrix resin. The fully cured structural part is then removed from the tool. This process has the potential for being cost effective and for producing very large complex structures.

Thermal Expansion Molding

Thermal expansion molding takes advantage of the "memory" characteristic of high temperature closed cell foam (see Figure 8). Polyamide foam is heated up to the softening point and compressed to a thickness on the order of 75 percent of its' original room temperature thickness. The compressed foam is then cooled down in the deflected state under pressure and removed from the press. Composite prepreg is laid up over the compressed foam surface and is inserted into the tooling cavity for the desired shape. Upon reheating, the foam expands which moves the composite material against the tool surface. Once the final cure temperature is reached the foam exerts sufficient pressure to consolidate the composite prepreg. The assembly is held at these conditions until the composite is fully cured.

Reaction Injection Molding

Reaction Injection Molding (RIM) is used to fabricate a wide variety of large self-supporting moldings for automotive and aircraft structures (see Figure 9). Polyurethane precursor materials are routinely employed for the fabrication of RIM parts. Automobile dashes, bumpers, removable hardtops, and splash aprons are made by this technique. Polyurethane resin and hardener are mixed and pumped under pressure (up to 20,000 psi) into the tool cavity after which it begins to rapidly foam and fill the cavity. The chemical reaction and cure time for this material system and process is on the order of 10 seconds, which makes this a very fast production process. Several million items are fabricated on an annual basis by this method.

Polymer Cure Cycles

In each of the aforementioned processes the polymer materials undergo either a chemical or physical change which renders the composite prepreg material a finished structural part. For thermoset resin matrix composite materials the prepreg is converted to a rigid structure after a polymer chemical reaction which can be initiated by the application of heat, radiation, catalysts, or other initiators. A typical cure cycle for epoxy materials is shown in Figure 10. Thermoplastic matrix materials are "cured" or converted to finished products by the physical shaping of the starting composite under heat and pressure. These materials have the capability to be reshaped in a follow-up processing cycle, therefore thermoforming is considered reversible fabrication process.

CONCLUSIONS

Advanced composite materials are being utilized to fabricate a wide variety of structures requiring high performance in specific environments. Applications for these materials are limited only by the imagination of designing engineers due to the versatility afforded by the vast array of combinations of fibers and resins available. The growth of this industry exceeds the growth of most industries in the USA. Applications for advanced composite materials range from tennis rackets to space shuttle components. The need for lightweight, high performance structures will continue to grow in all major manufacturing sectors. This is driven largely by increasing demands for conservation of resources and reduction of life cycle costs for consumable products. This growth pattern is expected to continue for an unlimited time in the future.

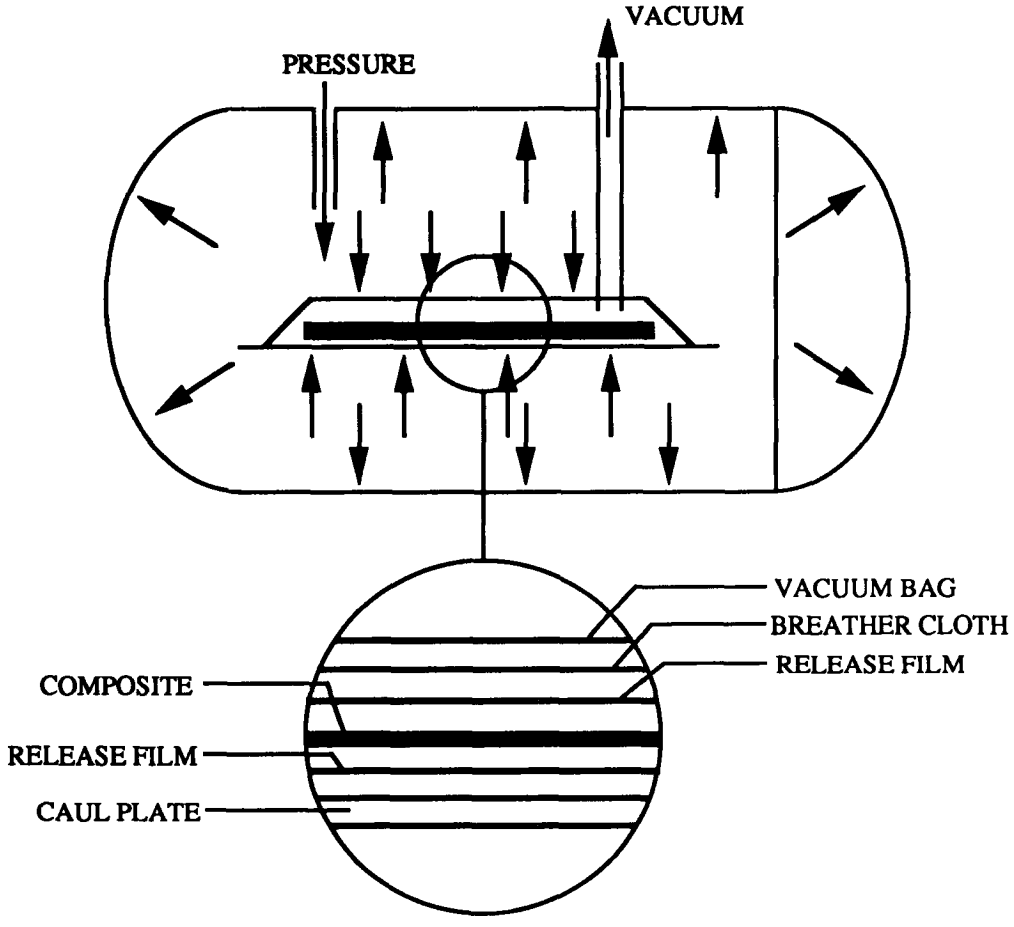


FIGURE 4. AUTOCLAVE AND VACUUM BAG SCHEMATIC

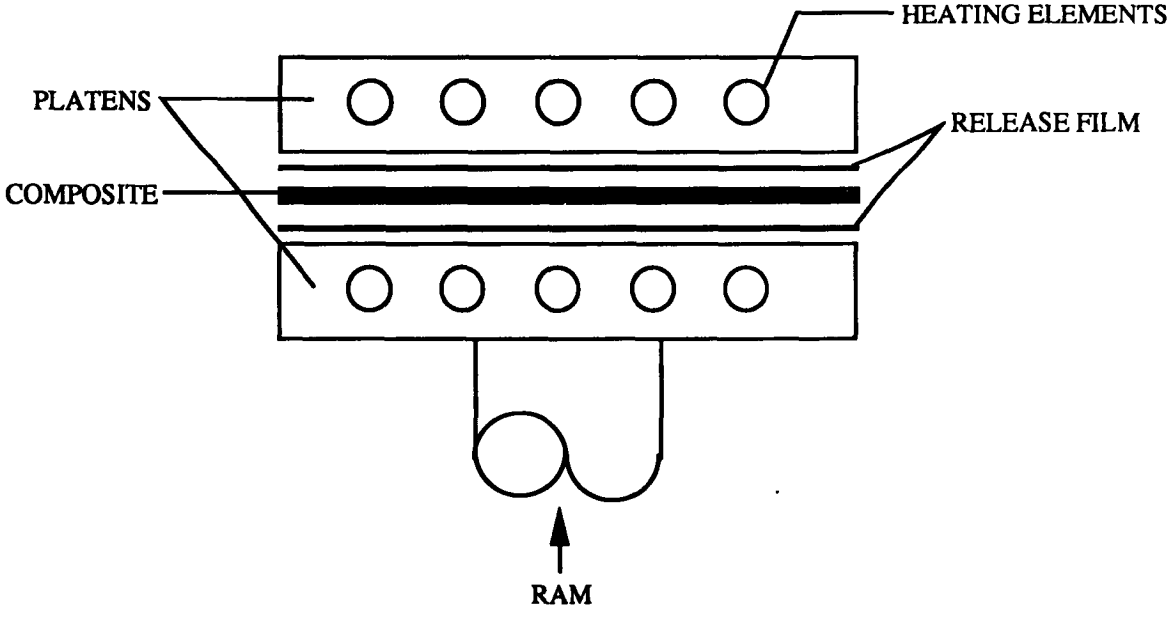


FIGURE 5. HEATED PLATEN PRESS

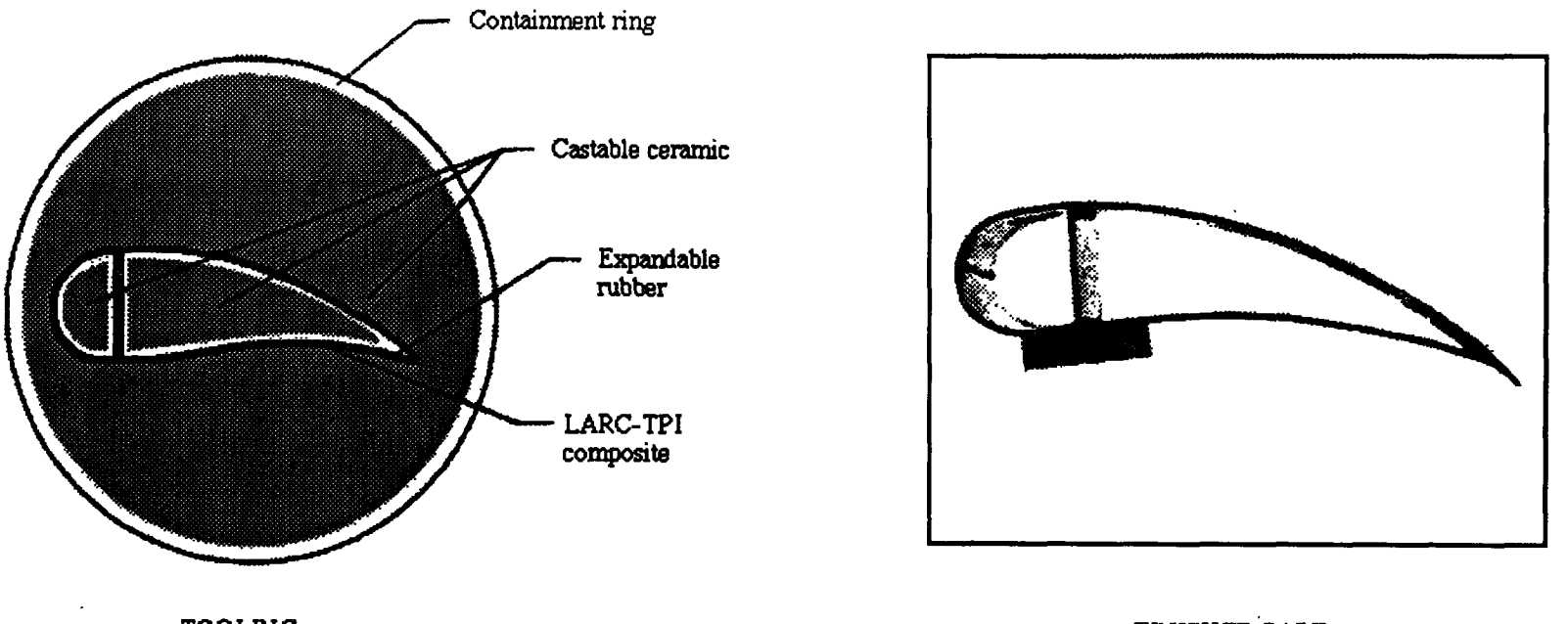


FIGURE 6. TRAPPED RUBBER MOLDING OF AIRFOIL

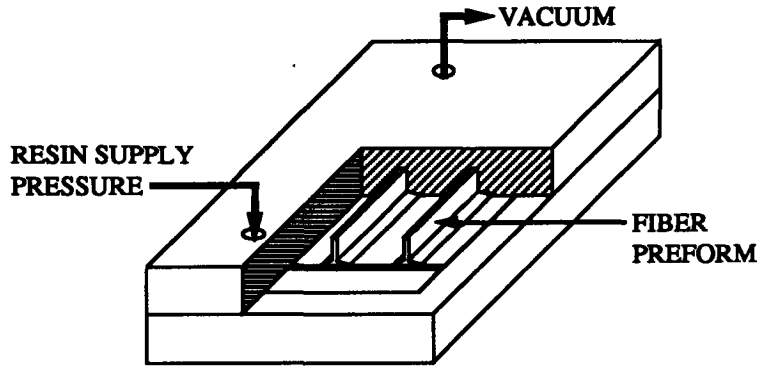


FIGURE 7. RESIN TRANSFER MOLDING

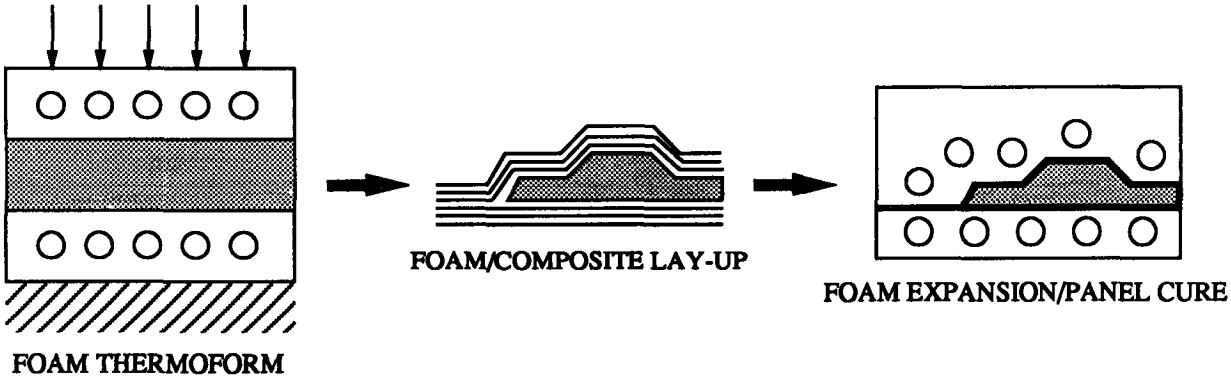


FIGURE 8. THERMAL EXPANSION MOLDING

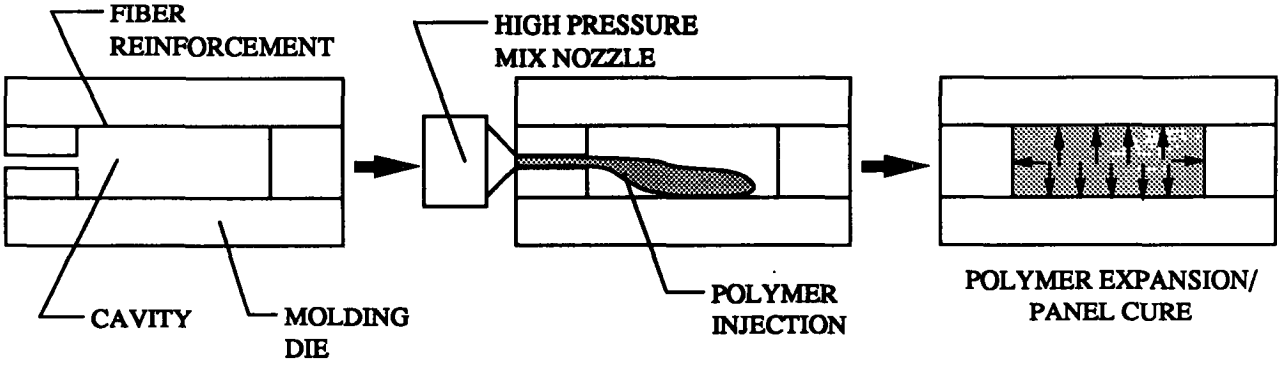


FIGURE 9. REACTION INJECTION MOLDING

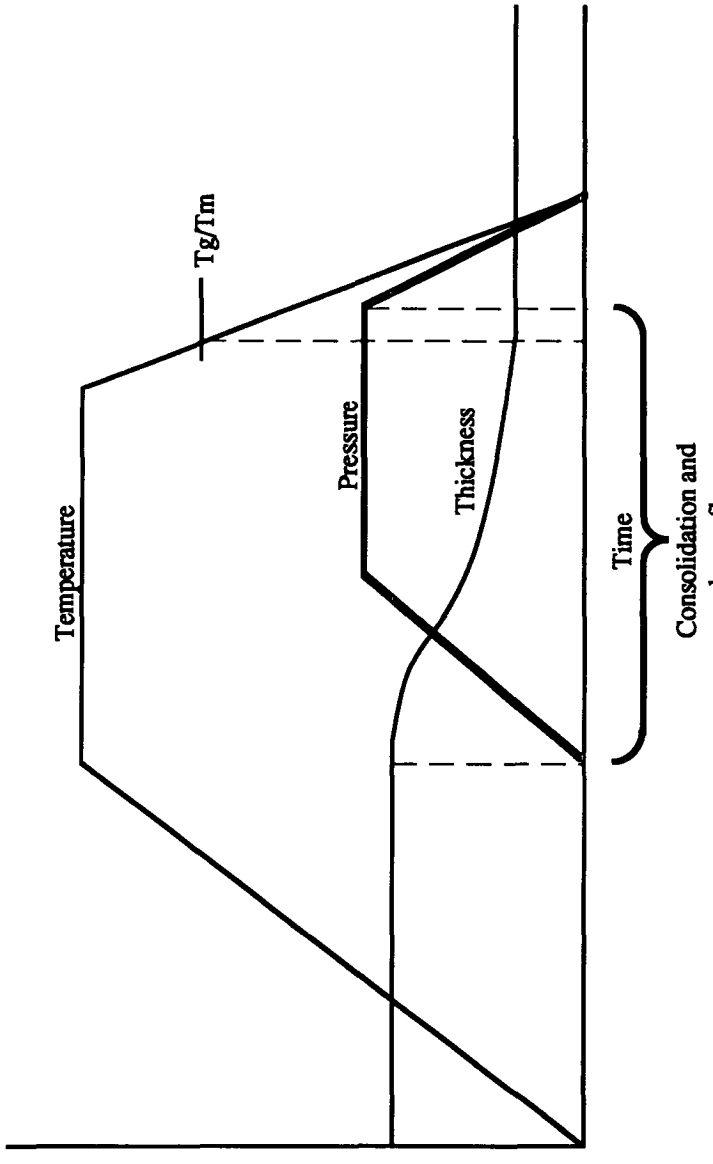
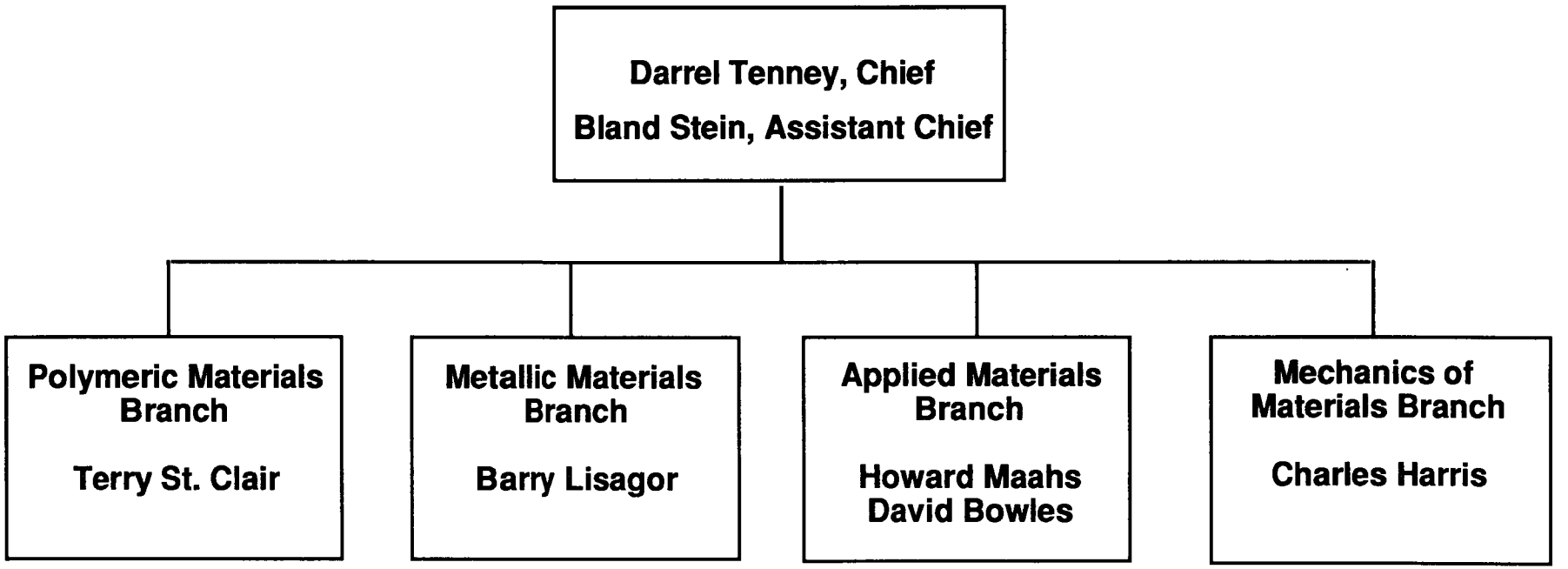


FIGURE 10. TYPICAL CURE PROFILE

MATERIALS DIVISION



- High-performance polymers
- Polymer charact.
- Tough composites

- Light alloy MMC development
- Innovative metals processing
- High temp. thin gage metallics

- Environmental effects
- Thermomechan. stability
- Carbon-carbon
- Advanced composite material concepts

- PMC damage tolerance
- Micromechanics
- TMF of MMC
- Fat. & Frac. of metals
- Constitutive models

N92-22678

RTM: COST-EFFECTIVE PROCESSING OF COMPOSITE STRUCTURES

Greg Hasko
Materials Research Engineer
Lockheed Engineering & Sciences Co.
Langley Program Office
144 Research Drive
Hampton, VA 23666
(804) 864-3093

H. Benson Dexter
Materials Research Engineer
NASA Langley Research Center
Mail Stop 188B
Hampton, VA 23665-5225
(804) 864-3094

ABSTRACT

Resin transfer molding (RTM) is a promising method for cost-effective fabrication of high-strength, low-weight composite structures from textile preforms. In this process, dry fibers are placed in a mold, resin is introduced either by vacuum infusion or pressure, and the part is cured. RTM has been used in many industries, including automotive, recreation, and aerospace. Each of the industries has different requirements of material strength, weight, reliability, environmental resistance, cost, and production rate. These requirements drive the selection of fibers and resin, fiber volume fractions, fiber orientations, mold design, and processing equipment. NASA Langley is sponsoring research to apply RTM to primary aircraft structures such as wings and fuselages, which require high strength and stiffness at low density, and are produced in relatively small quantities. However, there is a continuing need to reduce the cost of fabricating advanced composite structures.

This paper discusses the material requirements of various industries, methods of orienting and distributing fibers, mold configurations, and processing parameters. Processing and material parameters such as resin viscosity, preform compaction and permeability, and tool design concepts are discussed. Experimental methods to measure preform compaction and permeability are presented. Analytical methods that predict resin infiltration and cure are discussed. Mechanical properties and potential applications of selected textile material forms are also mentioned.

INTRODUCTION

Resin transfer molding (RTM) is a method of combining dry fibrous preforms with matrix resins in a mold to yield structural components. It has been used for making products over the last few decades in several industries. Parts are made with this process that often require little or no machining. The important relationships between end user (customer) requirements and component design and fabrication are depicted schematically in figure 1. There are significant differences in constituent materials and process parameters, depending on factors shown in figure 1. Overall selection of constituent materials is driven by the customer's performance needs, production volume, and cost sensitivity. Specific performance criteria are then defined and the design engineer can bracket the material selections. The type of fiber, its volume fraction, and orientation are selected by considerations of strength, stiffness, weight, and cost. Selection of resin is guided by service environment, lifetime requirements, reliability, and cost of the final part. The process engineer specifies temperature and pressure cycles and the equipment used to control them in order to produce repeatable parts within the allowable material processing windows. This paper discusses aterial and process selection for RTM, and analytical methods that help ensure cost-effective product development and production.

MATERIAL AND PROCESS SELECTION

Material Properties

Composite materials offer a wide range of material properties. This is illustrated in figure 2, which shows approximate specific strengths and stiffnesses of fiber composites and metals [1]. In comparison with composite materials, typical engineering metals have very narrow strength and stiffness ranges. Fiber composites have a considerably greater envelope in both strength and stiffness. Because of large differences in properties for fibers and resins, properties of composites can vary by a factor of 10 or more.

Composites allow tailoring of properties to the directions of applied loads, as shown in figure 3 [2, 3]. For a beam requiring stiffness in only one direction, a unidirectional graphite fiber reinforcement would be the stiffest and lightest. Skin panels of typical automotive or aerospace vehicles usually require uniform properties in all directions, so a multi-angle graphite laminate or a randomly oriented chopped glass composition should be considered. Aircraft wing skins are made of multidirectional graphite laminates, whereas automotive exterior panels use chopped glass molding compounds. Chopped glass is less structurally efficient, but less expensive than continuous graphite fiber reinforced composites because of lower fiber cost and simpler fiber deposition method.

The environmental performance of composite materials is strongly influenced by the nature of the resin. As shown in figure 4, polyester resin composites can lose 50 percent of their room temperature strength at 250°F. Epoxy resins can lose this same amount by 300°F, and bismaleimide resins by about 450°F [2, 3].

The weak link in a laminated composite is the interface between layers. Impacts from dropped tools or other items can cause delamination. Damage resistance of a composite is influenced by both fiber and resin selection. Figure 5 shows how strength retention after impact is increased by knitting and stitching the layers of reinforcement together and by using a toughened resin [4, 5]. In general, higher performance resins are more expensive and more difficult to process.

Cost Issues

Figure 6 shows the cost trends of currently used composite materials. Metals and low fiber volume glass composites used in the automotive industry are under $2 a pound [2]. There is a much broader range of high-performance fiber and resin costs, typically over $20 a pound for aerospace applications. Depending on stiffness and toughness requirements, the fiber or the resin may be the more expensive constituent. Cost is also process dependent: for RTM processing, the reinforcement and the resin are procured as separate components, which generally is the lowest cost form. For autoclave processing from prepreg tape, these constituents must first be combined by a prepregger prior to procurement by the part fabricator, adding another processing step and cost. More critical structures require greater control of fiber content and orientation, which adds cost.

Selection of processing methods to arrive at a given composition and orientation is guided by the cost of equipment and the volume of parts to be produced. Several thousand units of a given automobile frame are made yearly, whereas aircraft production may be as few as one hundred over several years. In automobile companies, production is based on market projections, whereas large aircraft manufacturers typically build only what is ordered. Trends of equipment cost amortized over production volume are shown in figure 7. Automotive composites manufacturers must invest in sufficient production capacity to meet assembly line rates. Aircraft companies, having lower production rates, require a lower investment. However, limited production volume can still cause high amortized unit costs.

Composite Processing Variations

For a given fiber content and orientation, many processes are available to fabricate the part. Originally, composites were mostly made by hand, laying down sheets of material or spraying chopped fibers on a mold. The laydown and spraying processes have been automated for greater control and efficiency. Continuous methods such as winding over a mandrel or pulling through a die are used. These methods place limits on the fiber orientation, especially if through-thickness reinforcement is desired. NASA is putting emphasis on textile methods to orient large quantities of fiber by rapid, automated processes. These methods include weaving and braiding of net shape structural elements with through-thickness fibers, as shown in figure 8.

An important cost factor in composite fabrication is how the matrix resin is introduced. One way is to apply resin before storing the raw material on a roll, as in prepreg. In order to prevent the resin from curing on the roll, it must be refrigerated. In pultrusion and filament winding, resin may be applied just before contact with the forming surface. Some resins are so viscous that fiber distortion or air entrapment are difficult to eliminate. With prepregs, it is impractical to introduce through-thickness fibers for damage tolerance, whereas RTM methods can allow processing for a wider variety of reinforcements and resins.

RESIN TRANSFER MOLDING

In usual RTM practice, dry fibers are oriented in a mold and resin is introduced from a reservoir. The reservoir may be in intimate contact with the perform or located several feet away in a heated tank connected by a heated hose to the mold. Resin flow direction can be through the thickness or in the plane of the preform. The type of mold depends on the compaction needed to obtain the final fiber volume fraction and on temperature requirements of the resin. Examples of these variations follow.

Tooling

Figure 9 shows a patented process in which flow into the mold occurs in three steps: from one inlet to a lengthwise conduit, then laterally across the surface area, then through the thickness of the preform [6]. A single-sided hard mold is used, with the other side consisting of layers of porous materials and a vacuum bag. Figure 10 shows a similar through-the-thickness flow in the preform, but the reservoir consists of a solid film of resin placed in the mold. The resin melts and flows into the preform under heat and pressure. The compaction pressure is also the resin infiltration pressure. Figure 11 shows the same process as figure 10 applied to a typical aircraft wing skin with stiffening members attached to the skin [7]. Heat and pressure are applied with an autoclave.

Figure 12 is a typical pressure injection mold with an external reservoir. The preform compaction pressure is applied independently of resin injection pressure. Figure 13 is a patented version of an in-plane pressure injection process for forming hollow parts. The inside surface is defined by a pressurized bladder [8].

Mathematical Process Models

Process parameters for a specific part have historically been based on experience. Mold design, heating method, flow path, and time/temperature/pressure cycles were determined by past experience and "build and break" trials. NASA Langley has been sponsoring research on reducing the need for a trial-and-error approach by applying a science-based understanding of the process. Success of the science-based approach depends on accurate data for the preform and resin processing behavior, and on mathematical models of the process.

The flow of resins through fiber performs can be modeled in a similar manner as the flow of a fluid through any porous media. Darcy's Law states that the flow rate (Q) is directly proportional to the pressure gradient $\Delta P/X$, area (A), and permeability (K); and inversely proportional to viscosity (μ):

$$Q=\left(\frac{K}{\mu}\right)\frac{\Delta P}{X}A \qquad (1)$$

The permeability of a preform is dependent on the fiber size, weave geometry, packing density, and direction of flow. The viscosity of the resin is dependent on its entire time/temperature history. These parameters can be measured and modeled in order to predict flow rate and pressure for any type of preform and any cure cycle. Currently, the preform and resin manufacturers do not supply the data needed to successfully model the RTM process. NASA Langley is supporting development of material characterization methods and process models.

Preform Compaction and Permeability

Preform compaction behavior, needed for input into process models, is measured in a fixture where a known compaction pressure is applied to a sample and the thickness is measured, figure 14. Fiber volume fraction is directly proportional to thickness but is indirectly related to compaction pressure. As shown in the figure, preforms in an ambient, uncompacted state range from 30 to 52 percent fiber volume. Vacuum bag pressure (14.7 psia) shifts this range to 35 to 55 percent. For automotive and recreational products, these values may be acceptable. However, high-performance aerospace components require high stiffness to weight ratios. Material selection is driven towards graphite composites of 60 percent fiber volume, which require 30 to 130 psi for proper compaction.

Permeability information for a preform is determined by compacting a sample of the material to a known fiber volume and pumping a fluid of known viscosity, figure 15. Flow rate and pressure drop are recorded and permeability is calculated using equation (1). This test is done at varying fiber volumes and in the three primary preform directions. Permeability has been found to vary by a factor of 100 when measured along and across a fiber bundle. Resin can flow much easier along a bundle than across it. In addition, as shown in figure 15, permeability can vary by a factor of 100,000 between a glass mat at 25 percent fiber volume and woven graphite at 70 percent fiber volume [9, 10, 11].

Resin Viscosity Behavior

Resin viscosity can vary by several orders of magnitude, depending on its chemistry, aging history, and temperature. Figure 16 shows viscosity behavior for three different epoxies (Hercules 3501-6, British Petroleum E905L, and 3M PR500), each at two selected temperatures.[1] In general, higher processing temperatures reduce the initial viscosity but cause the viscosity to increase at a faster rate as cure progresses. For small parts or high permeability preforms, a higher temperature will allow quick wet out. However, with large parts or low permeability preforms, the resin may gel prior to full impregnation if the mold is designed for in-plane flow. Success may be achieved with a through-thickness flow path.

Figure 17 shows an example of the processing window for a hot melt epoxy. This material melts as the mold heats up, reaches a minimum viscosity, then reacts and thickens. All of the flow into the preform must occur before the viscosity starts to increase.

Versatility of Textile Equipment

It is possible to produce preforms for both the automotive and aerospace industries using a common textile machine. For example, the braider shown in figure 18 can produce preforms for automobile bodies or for jet engine inlets. To successfully complete the process development cycle, both industries would benefit from a mathematical model of the RTM process to aid in mold design and process control.

RTM Computer Model

Under a NASA grant, Virginia Polytechnic Institute and State University has been developing the computerized model of the RTM process shown schematically in figure 19. Currently, the model simulates only one-dimensional flow occuring in a through-thickness process. The user selects mold design and process parameters, and the program calculates the variation in compaction, permeability, heat transfer, cure kinetics, and resin flow with time. The program calls on subroutines for specific preform and resin characteristics. With sufficient data, the program accurately models the behavior of the preform and resin inside the mold up to final cure of the part. The model is being extended to include two- and three- dimensional flow conditions. Computing requirements for the model are also shown in figure 19. Model predictions have been verified for several one-dimensional cases by fabricating panels in instrumented molds.

[1]The use of trademarks of names of manufacturers in this paper does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

ORIGINAL PAGE IS
OF POOR QUALITY

RTM Flow Sensing and Control

The College of William and Mary has developed a sensor system under a NASA grant to monitor resin parameters during composite processing. Termed Frequency-Dependent Electromagnetic Sensors (FDEMS), a single sensor can monitor the degree of wet out, viscosity, and cure state of a resin. Several sensors can be placed in an RTM mold to monitor many locations at once. An additional benefit of such a monitoring system is that sensor output can be used to control the process in real time, based on actual resin parameters, as opposed to indirect parameters such as mold surface temperature. This system can be used during process development or in production to control mold temperature, as shown in figure 20.

CONCLUSIONS

Composite materials offer the designer a wider range of properties and higher performance than metals. However, raw material and processing costs of composites are a barrier to their greater utilization in industry. The key to greater acceptance of composites lies in process innovations that result in a cost-effective product.

RTM processing allows fabrication of composite materials into components of complex shape with any desired fiber orientation and all but the most viscous resins. However, mold designs and process parameters can vary tremendously. Analytical tools are needed to identify the key variables leading to a cost-effective product and process. NASA's efforts in science-based understanding of RTM can guide mold design and process development and help control process variables in production. The same approach can be useful in an industrial environment, ranging from the automotive body panel made from chopped glass and polyester resin in an in-plane mold at room temperature, to an aircraft wing skin made of woven, stitched graphite and hot melt epoxy in a through-thickness process. In conjunction with near- net shape textile preforms, RTM shows great promise to be a cost-effective process for producing high-quality composite structures.

References

1. Jones, R. M: Mechanics of Composite Materials. Scripta Book Co., 1975, pp. 92.
2. Engineered Materials Handbook Volume 1 - Composites. ASM International, 1987, pp. 220, 379, 395, 411.
3. DOD/NASA Advanced Composites Design Guide. p.4.1.1, July 1987.
4. Dexter, H. B.; Hasko, G. H.; and Cano, R. J.: Characterization of Multiaxial Warp Knit Composites. Presented at the First NASA Advanced Composites Technology Conference, NASA CP-3104, 1990.
5. Dow, M. B.; and Smith, D. L.: Properties of Two Composite Materials Made of Toughened Epoxy Resin and High-Strain Graphite Fiber. NASA TP-2826, July 1988.
6. Seeman, W. H.: U.S. Patent #4,902,215.
7. Markus, A.; and Palmer, R.: Resin Transfer Molding for Advanced Composite Primary Structures. Presented at the First NASA Advanced Composites Technology Conference, NASA CP-3104, 1990.
8. Freeman, R. B.: U.S. Patent #4,911,876.
9. Kim, Y. R., et. al.: Resin Flow Through Fiber Reinforcements During Composite Processing. 22nd International SAMPE Technical Conference, November 1990.
10. Trevino, L., et. al.: Analysis of Resin Injection Molding in Molds with Preplaced Fiber Mats. 1: Permeability and Compressibility Measurements. Polymer Composites, Vol. 12, No. 1, February 1991.
11. Gutowski, T. G., et. al.: Consolidation Experiments for Laminated Composites. Journal of Composite Materials, July 1987.

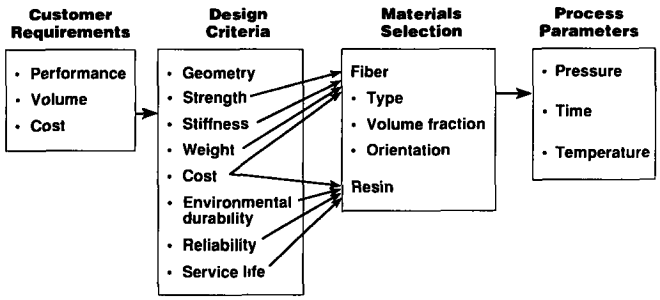


Figure 1. Material and Process Drivers

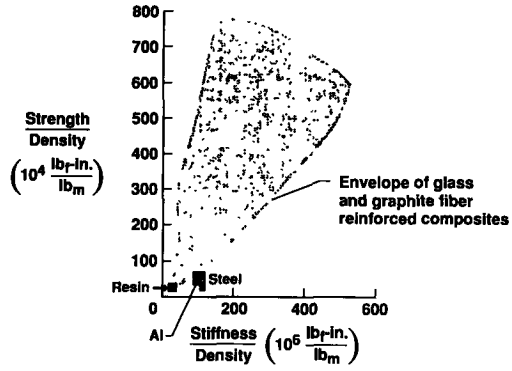


Figure 2. Specific Strength and Stiffness Advantages of Composite Materials

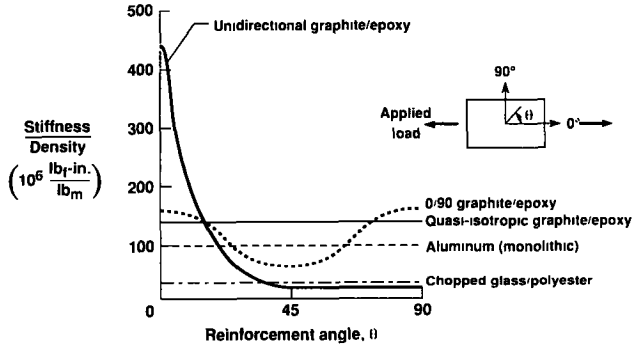


Figure 3. Specific Stiffness as a Function of Reinforcement Orientation

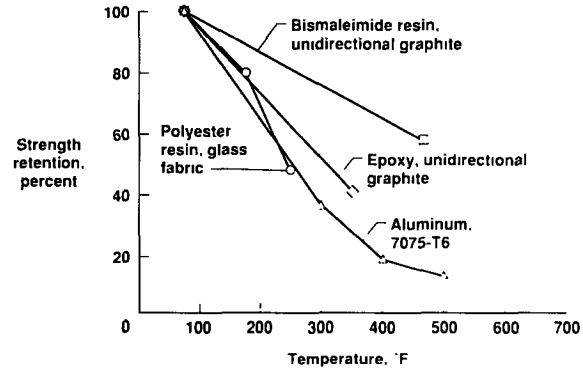


Figure 4. Effect of Temperature on Strength Retention of Materials

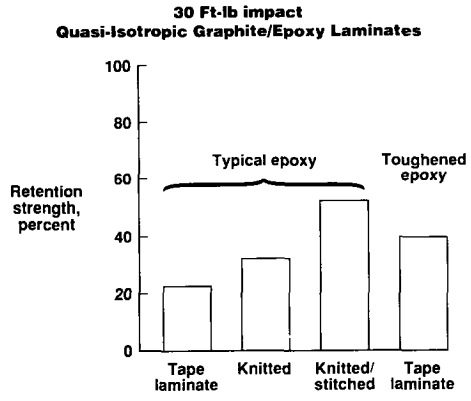


Figure 5. Compression after Impact Strength

Processing and Scrap Costs Omitted

Material	Cost, $/lb	End use
Steel	0.38	Automotive
Aluminum	1.44	Automotive
Glass-filled thermoplastic	1.70	Recreation
Low performance graphite/epoxy (sum of constituents for RTM processing)	17 00	Recreation
High performance graphite/epoxy (sum of constituents for RTM processing)	33 0	Aerospace
High performance graphite/epoxy (prepreg cost)	40.00	Aerospace

Figure 6. Raw Material Cost Comparison

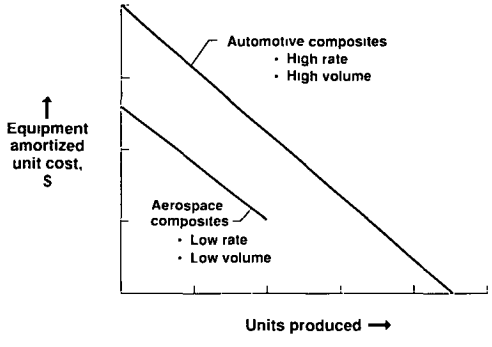


Figure 7. Relative Amortized Equipment Cost Per Unit Produced

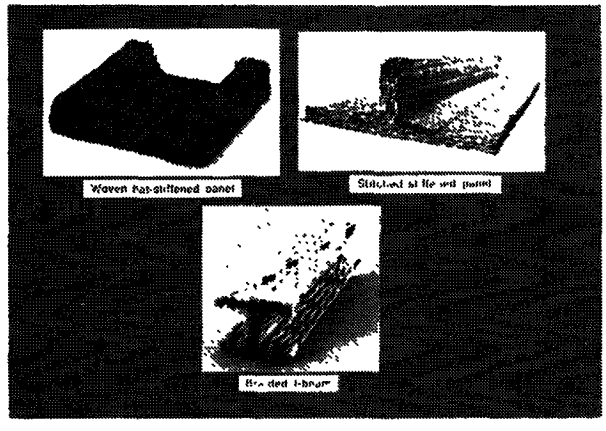


Figure 8. Net-Shaped Fiber Preforms

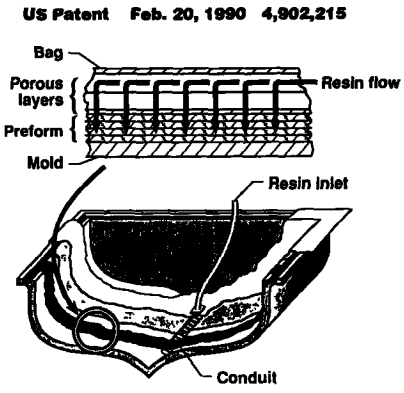


Figure 9. Liquid RTM, Bag Mold, Through-The-Thickness Flow

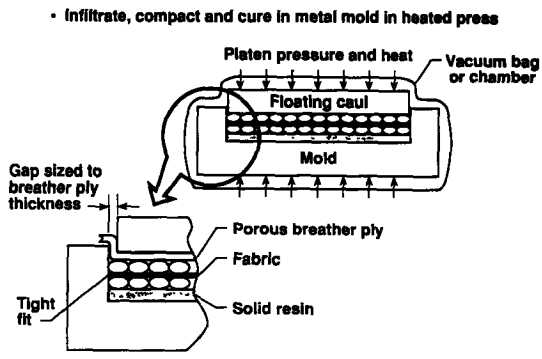


Figure 10. One-Step Through Thickness Vacuum/Pressure Process

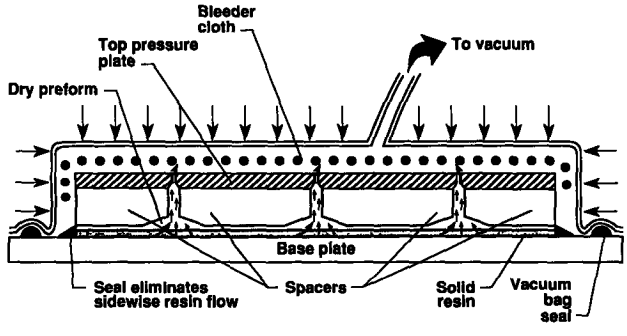


Figure 11. Resin Film Infusion of Stiffened Panel

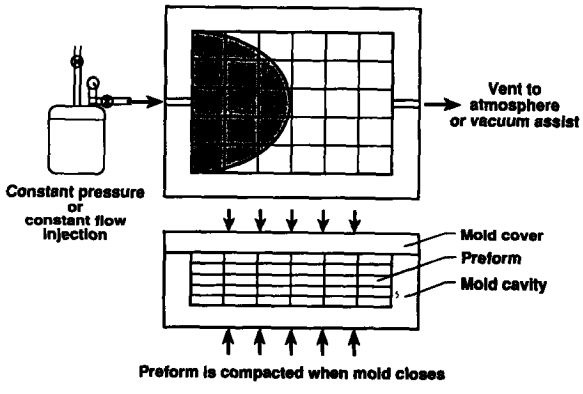


Figure 12. Pressure Injection

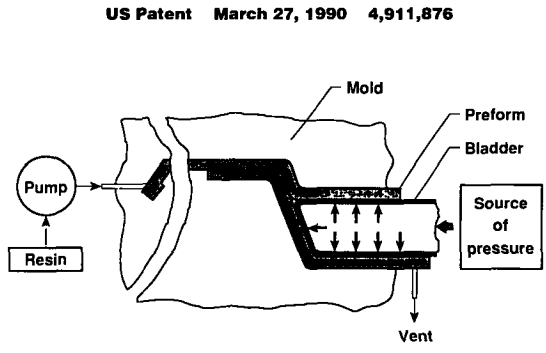


Figure 13. Liquid RTM, Inflatable Bag Mold, In-Plane Flow

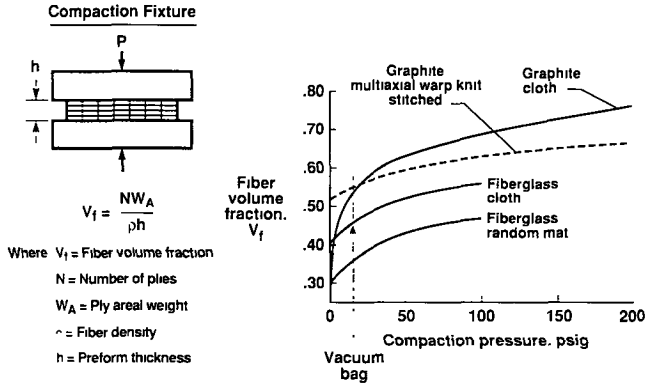


Figure 14. Preform Compaction Behavior

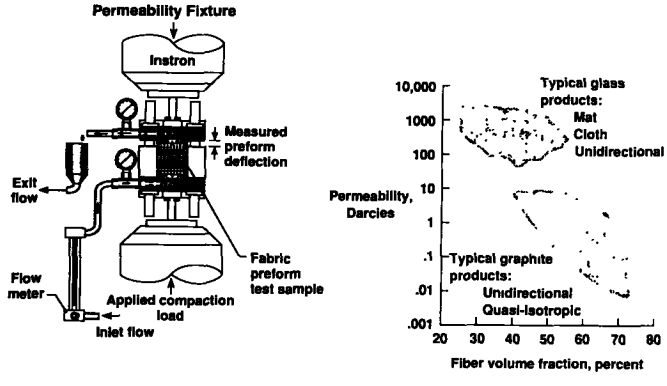


Figure 15. Permeability Comparison

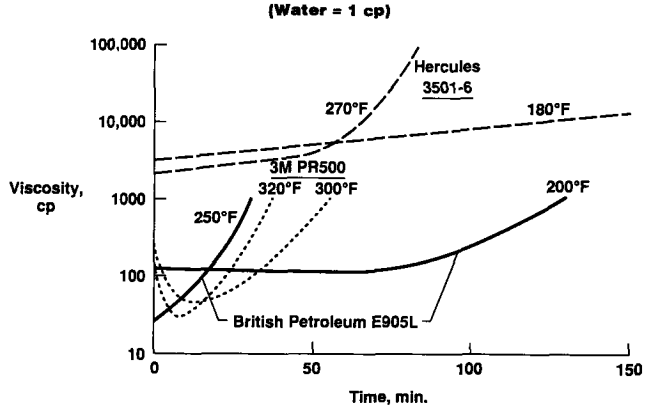


Figure 16. Effect of Time and Temperature on Viscosity of Epoxy Resins

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

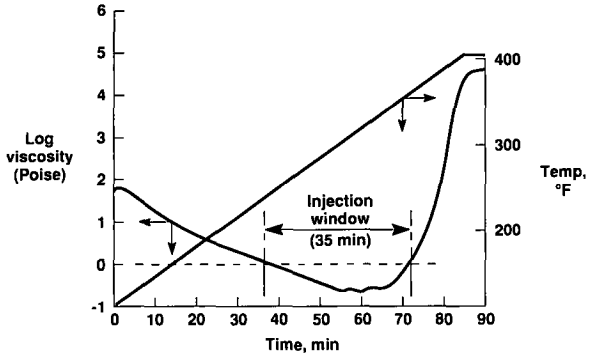


Figure 17. Viscosity Profile for PR500 Epoxy

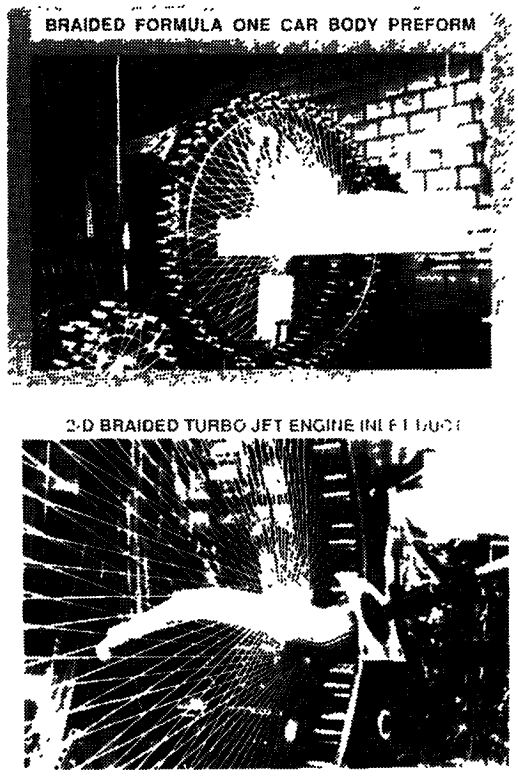


Figure 18. Braiding Preforms for Automotive and Aerospace Use

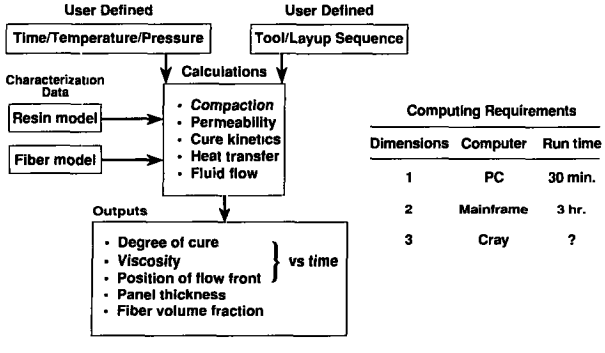


Figure 19. RTM Computer Model

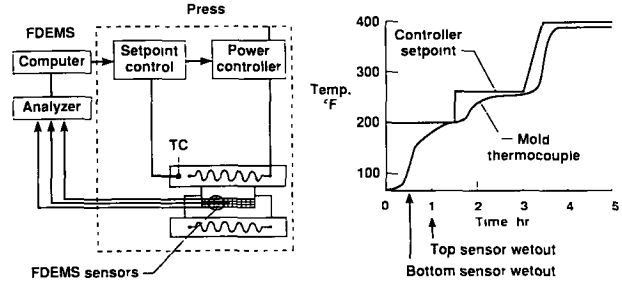


Figure 20. Prototype FDEMS Expert Cure System

N92-22679

A LOW COST METHOD OF TESTING COMPRESSION-AFTER-IMPACT STRENGTH OF COMPOSITE LAMINATES

Alan T. Nettles
NASA Marshall Space Flight Center
Polymers and Composites Branch (EH33)
MSFC, AL 35812

ABSTRACT

A method has been devised to test the compression strength of composite laminate specimens that are much thinner and wider than other tests require. The specimen can be up to 7.62 cm (3 in.) wide and as thin as 1.02 mm (.04 in.). The best features of the Illinois Institute of Technology Research Institute (IITRI) fixture are combined with an anti-buckling jig developed and used at the University of Dayton Research Institute to obtain a method of compression testing thin, wide test coupons on any 20 kip (or larger) loading frame. Up to 83% less composite material is needed for the test coupons compared to the most commonly used compression-after-impact (CAI) tests, which call for 48 ply thick (~ 6.12 mm) test coupons. Another advantage of the new method is that composite coupons of the exact lay-up and thickness of production parts can be tested for CAI strength, thus yielding more meaningful results. This new method was used to compression test 8 and 16 ply laminates of T300/934 carbon/epoxy. These results were compared to those obtained using ASTM standard D 3410-87 (Celanese compression test). CAI testing was performed on IM6/3501-6, IM7/SP500 and IM7/F3900. The new test method and associated fixture work well and will be a valuable asset to MSFC's composite materials damage tolerance program.

INTRODUCTION

Since the most critical damage tolerance feature of structural composite materials is their ability to carry a compressive load after damage, a simple, inexpensive method of testing this characteristic needs to be established. The two most commonly used methods, the Boeing and NASA CAI tests, both call for the use of 48 ply thick specimens. There are four big disadvantages to using such a thick specimen. 1) Most CAI testing is on new and experimental materials which are either expensive, in limited supply, or both. It would save time and money if less material were needed. 2) Foreign object impact characteristics are much different on a 48 ply specimen than on 16 or 8 ply specimens, regardless of boundary conditions. Since most functional parts are commonly in the vicinity of 16 plies in thickness, it would yield more meaningful data to test the actual lay-up sequence that the final product calls for. 3) A large amount of load is needed to cause failure in the Boeing and NASA 48 ply specimens. A much smaller, less expensive load frame can be used if 16 ply (or less) laminates are used. 4) Approximately 334 J (244 ft-lbs) of elastic energy can be stored in the 48 ply test specimens, (all of which is released upon specimen rupture), whereas the 16 ply specimens will only store about one-fourth as much energy, making for a safer test.

In order for impact damage to be accurately characterized by CAI testing, a long, wide gage length is needed to entirely contain the impact damage. This requirement, coupled with the desire to test specimens much thinner than 48 plies calls for a method to prevent global buckling of the compression specimen. Ryder and Black (1) wrote on compression testing large gage length specimens in 1977. They used a face-supporting fixture based on ASTM Standard Test for Compressive Properties of Rigid Plastics (D 695-69). This fixture made contact with the entire gage length surface of 140 mm (5.5 in.) long, 22.2 mm (.874 in.) wide, 16 ply specimens, tabbed and shear-loaded at one end and end-loaded at the other. Clark and Lisagor (2) introduced a face-supported fixture in 1981 that tested specimens as thin as 8 plies and up to 50 mm (2 in.) wide and 152 mm (6 in.) long. These specimens were tabbed at each end and tested in a hydraulic grip system. The anti-buckling jig was made up of inner and outer platens on each side of the specimen. The Boeing Open Hole Compression Test Standard (BSS 7260) is also a face-supporting compression test fixture (end-loaded).

Sjoblom and Hwang (3) of the University of Dayton Research Institute introduced a very simple method of supporting the gage length of a thin, wide compression test coupon in order to prevent global buckling of the specimen. This technique utilized two metal plates that would sandwich the test specimen along all but 1.9 mm (.075 in.) of the gage length. These plates were secured with just enough pressure to prevent the plates from freely moving on the specimen. In order to accommodate CAI specimens, holes were machined into the center of the plates to allow room for the protruding damage zone. A MTS hydraulic grip system was used to secure the specimens for compression loading. Since availability to hydraulic grips may be limited, Marshall Space Flight Center has

developed a CAI fixture that can be used on any loading frame of 90 kN (20,000 lbs) capacity or larger. This fixture is a modified IITRI test apparatus that can accomodate specimens up to 76.2 mm (3 in.) wide. A face support system much like that used by Sjoblom and Hwang is used to prevent global buckling of the specimen.

TEST FIXTURE

A drawing of the test fixture labeling its components is shown in Figure 1. A photograph of the loaded test fixture and a view of the clamping wedges and load-alignment block is presented in Figures 2 and 3. Figures 4, 5 and 6 contain detailed drawings of the components of the test fixture. The entire fixture is fabricated of stainless steel (except for the anti-buckling faceplates) and measures approximately 28 cm (11 in.) in height when loaded with a specimen. The University of Dayton Research Institute's faceplate design was modified by increasing the cutout area to accommodate the damage zone which was often too large. A rectangular shape of 5.1 X 2.5 cm (2 X 1 in.) was utilized since the damage zone tended to protrude out lengthwise to the fibers in the outer ply. Thus all testing must be performed with the outer plies in the 0° direction (vertically). Like the Celanese fixture, the modified IITRI fixture was fitted with an outer sleeve to aid in proper alignment of the fixture and also to act as a protective shield should the fixture fail. In addition, four alignment rods were used between the upper and lower load-alignment blocks instead of two, as on the IITRI fixture. The entire fixture weighs in at a hefty 34 kg (74 lbs) but is set up in the loading frame by sections so it never needs to be lifted as one unit. All moving parts are greased to allow smooth movement. The anti-buckling faceplates were bolted onto the gage length of the specimen with just enough pressure so the faceplates would not move freely on the specimen. The inner surfaces of the faceplates were sprayed with a Teflon coating before each test to assure that fiction between the specimen and faceplate would not be a factor. The faceplates were machined from 16.8 mm (.66 in) thick aluminum to prevent any bending like that reported by Clark and Lisagor (2).

SPECIMEN PREPARATION

The specimen dimensions were kept the same as those used by Sjoblom and Hwang (3) and are given in Figure 7. The fiberglass end tabs were processed so one side would contain a crisscross pattern to allow the wedge grips to better "bite" into the specimen. This was achieved by using a Teflon coated, woven fiberglass cloth as a peel ply on one of the sides. The other side was smooth to aid in the adhesion of the glass/epoxy tabs to the carbon/epoxy specimen. It is important to note that no adhesive should run out from under the glass tabs and onto the gage length of the specimen since the faceplates must fit properly into this region. This can be accomplished by using flashbreaker tape or a similar non-stick substance at the area were the glass tabs edges meet the carbon/epoxy, or by using just the right amount of adhesive between the glass tabs and the specimen so no excess is produced. At MSFC it was found that a 19 mm (.75 in) wide strip of Cyanamid's FM 300 film adhesive, placed at the top edge of the glass tabs would produce acceptable specimens (see Figure 7.).

IMPACT TESTING

A Dynatup 8200 instrumented drop weight apparatus was used in this study to inflict impact damage on the carbon/epoxy specimens. The falling crosshead was outfitted with a 1.27 cm (.50 in.) diameter tup and had a mass of 1.77 kg (3.9 lbs). The specimens were impacted at their geometric centers and held fast by a pneumatic clamping device over a 6.35 cm (2.5 in.) diameter hole. Just about any specimen support and impact device can be used, as long as the damage zone is not so large as to cover the entire specimen width.

TEST RESULTS

Test of Faceplate Stiffness Criticality

Since it has been reported that the stiffness of the specimen face supporting jig can be a critical parameter (2,3), a series of compression tests were performed on 16 ply, undamaged laminates of T300/934 with three different thicknesses of aluminum faceplates, and a 6.1 mm (.24 in.) thick stainless steel faceplate. The thinnest aluminum faceplate was 6.1 mm (.24 in.) thick and gave an average breaking stress of 310 MPa (45,000 P.S.I.). The next aluminum faceplate tested was 16.8 mm (.66 in.) in thickness and gave an average breaking stress of 482 MPa (70,000 P.S.I.). The thickest aluminum plate measured 25.4 mm (1.0 in.) and also gave a breaking stress of 482 MPa (70,000 P.S.I.). The steel faceplate gave a value of 455 MPa (66,000 P.S.I.) Thus it was concluded that the 16.8mm (.66 in.) aluminum faceplates were robust enough not to deflect significantly enough to affect the outcome of the tests and were utilized for the remainder of the test program.

Comparison With Celanese Fixture

Compression tests were carried out on undamaged specimens of 16 ply $(0, +45, 90, -45)_{S2}$ T300/934 carbon/epoxy utilizing ASTM Test Standard D3410 (Celanese compression) and the new fixture presented in this paper. A total of 26 Celanese tests and 16 tests with the new fixture were performed. The average compression breaking stress for the Celanese test specimens was 434 MPa (63,000 P.S.I.) with a standard deviation of 55 MPa (8,000 P.S.I.). The new fixture gave an average compression breaking stress of 482 MPa (70,000 P.S.I.) with a standard deviation of 41 MPa (6,000 P.S.I.). Although the values are close (within 11 %), previous studies (2) have shown that a face-supported compression test specimen yields values slightly *lower* than a short gage length test. However, this study utilized very robust faceplates which ensured no out of plane stresses in the specimen. The low standard deviation seen with the new fixture suggests that the friction between the faceplates and the specimen is negligible since the torque on the bolts in the faceplates was a very arbitrary value that was not measured with any instruments. In addition, some tests were performed without any Teflon lubricant on the faceplates and values were obtained which were not significantly different than when the Teflon was used. Edge views of broken specimens from each type of test are presented in Figure 8. These breaks are typical of specimens that have undergone extensive interlaminar failure between dissimilar oriented plies, (see references 1,2).

CAI Testing

Compression-after-impact tests were performed on three different materials, IM6/3501-6, a standard early generation carbon/epoxy system and two new toughened systems, IM7/SP500 and IM7/F3900. A 16 ply quasi-isotropic layup configuration was used, $(0, +45, 90, -45)_{S2}$. A large range of impact energies was used to better understand and compare the materials. As expected, the two new, toughened systems could carry much more load at a given impact energy level than the old generation IM6/3501-6. A plot of impact energy versus residual strength is given in Figure 9. Figure 10. normalizes this data by laminate thickness.

Compression Testing of 8 Ply Specimens

A total of six undamaged and two impact damaged 8 ply quasi-isotropic specimens of T300/934 were tested in the new fixture. The average undamaged strength was 407 MPa (59,000 P.S.I.), much lower than the 16 ply specimen's average value of 482 MPa (70,000 P.S.I.). The impact damaged specimens, hit with 1.2 J (.88 ft-lbs) of incident impact energy, failed at 282 MPa (41,000 P.S.I.) and 276 MPa (40,000 P.S.I.). Sixteen ply quasi-isotropic specimens impacted at 1.2 and 2.4 J (.88 and 1.76 ft-lbs) had CAI strengths of 447 and 319 MPa (65,000 and 46,000 P.S.I.) respectively. While the residual strength of impact damaged 8 ply specimens are difficult to compare with thicker specimens,mostly due to the different damage mechanisms involved during impact, the fixture does cause failure of the 8 ply specimens at the impact damage zone. Undamaged 8 ply coupons gave values significantly lower than the 16 ply specimens. Thus 8 ply specimens that have been damaged severely enough to cause a drop in strength can be tested with the new fixture, but it is not recommended to test for virgin strength.

CHARACTERISTICS OF THE NEW FIXTURE

As the test program evolved, some methods to better test the compression coupons were discovered.

Tabbing the Specimen

As mentioned earlier in this paper, the tabs to be bonded to the test coupon needed to be applied so that no flashing of adhesive would occur in the gage length of the specimen so as to not interfere with the anti-buckling faceplates. It was found that on occasion the wedge grips would not "bite" into the tabs and would simply slide down. This problem was solve by "roughening" the outer surface of the fiberglass tabs during processing by using a Teflon coated, woven glass fabric as a peel ply on one side of the glass/epoxy laminate to be used for tab material. In rare instances, one of the tabs would debond from the specimen. This was usually the result of the adhesive material between the carbon/epoxy specimen and the tabs not being fully cured (as indicated by color of the adhesive). If a tab does debond it can be picked up on the load curve as a sudden, but not large drop in force. If this occurs the test should be stopped immediately and the specimen checked since extreme uneven loading would occur, possibly damaging the fixture. In this study a tab debond occurred and the test was not stopped and the result was four bent pins between the clamping wedges. Fortunately these were easily replaced.

Placing Faceplates on Specimen

Although it was determined that friction between the faceplates and specimen was negligible, a Teflon spray was applied to the faceplates prior to testing and wiped clean after the test had been performed to assure no catching of the specimen by the faceplate. Teflon tape was tried but was found to be too easily damaged, thus the use of spray. The four bolts and nuts that secured the faceplates to the specimen were finger tightened in a crisscross pattern with great care being taken to ensure that the faceplates were completely flat against the surface of the carbon/epoxy specimen. Calipers were used to make sure that there was an even gap between the two faceplates around their perimeters. The faceplates were mounted on the specimen before loading into the fixture.

Loading of Specimens

The most important aspect of loading the specimen into the fixture is to make certain that the specimen's length is perpendicular to the loading platens. This is not very difficult since the diamond pattern on the wedge grips is the same width of the specimen. Thus if the tabs are in contact with the diamond pattern without any overhang, the specimen is assured of proper alignment. In addition, if the specimens were tabbed and machined properly, any tab should be parallel to the clamping wedge that grips it.

The specimen, with faceplate, is then loaded into the bottom clamping wedge and load-alignment block which have previously been set on the bottom loading platen on the test frame. The four alignment rods are then placed in the bottom load-alignment block. The next step is to place the upper clamping wedges on the specimen and allow them to rest on the top surface of the faceplates. The upper load-alignment block is then placed on the four alignment rods and allowed to slide down over the clamping wedges. The wedges are then lifted slightly as is the loading block until the grips are at the point desired on the tabs. A little pressure pushing the clamping wedges up into the load-alignment block will lock the upper load alignment block in place. The outer sleeve is then placed over the entire fixture and the top loading platen is brought down to the top surface of the upper load-alignment block. Testing is now ready to begin.

Maintenance of Fixture

All moving parts are cleaned and regreased after 6-8 tests have been performed. The fixture is carefully examined for any galling or pitting of the metal parts. The most critical area of the fixture is the mating of the tapered surface of the clamping wedges to the inner tapered surface of the load-alignment block. These surfaces must always be clean and well greased.

CONCLUSIONS

The CAI fixture presented in this paper has been used successfully at MSFC for damage tolerance testing of composite materials. The device allows a small (20 kip) load frame to be utilized, thus saving time and cost by not having to test outside the Polymers and Composites Branch. In addition, much less material is needed to fabricate a CAI test specimen which also saves time and money. Furthermore, more specimens can be fabricated thus allowing a larger range of impact energies to be tested.

REFERENCES

1. Ryder, J.T. and Black, E.D., "Compression Testing of Large Gage Length Composite Coupons," Composite Materials: Testing and design (Fourth Conference)' ASTM STP 617, 1977, pp. 170-189.

2. Clark, R.K. and Lisagor, W.B., "Compression Testing of Graphite/Epoxy Composite Materials," Test Methods and Design Allowables for Fibrous Composites, ASTM STP 734, 1981, pp. 34-53.

3. Sjoblom, P. and Hwang, B., "Compression-After-Impact: The $5,000 Data Point!," Proceedings of the 34th International SAMPE Symposium, Reno, NV, 1989, pp. 1411-1421.

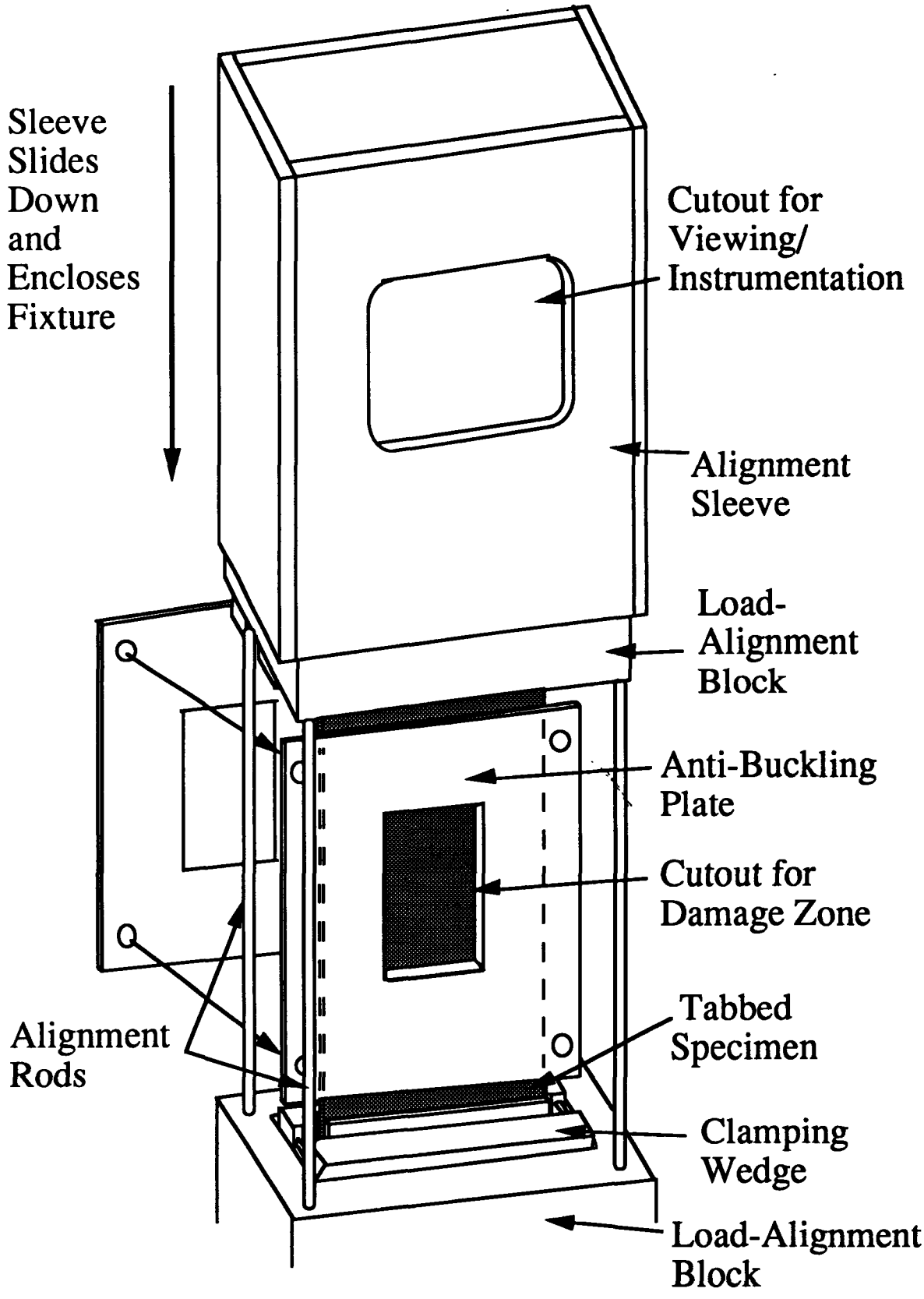


Figure 1. New Compression-After-Impact Fixture

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

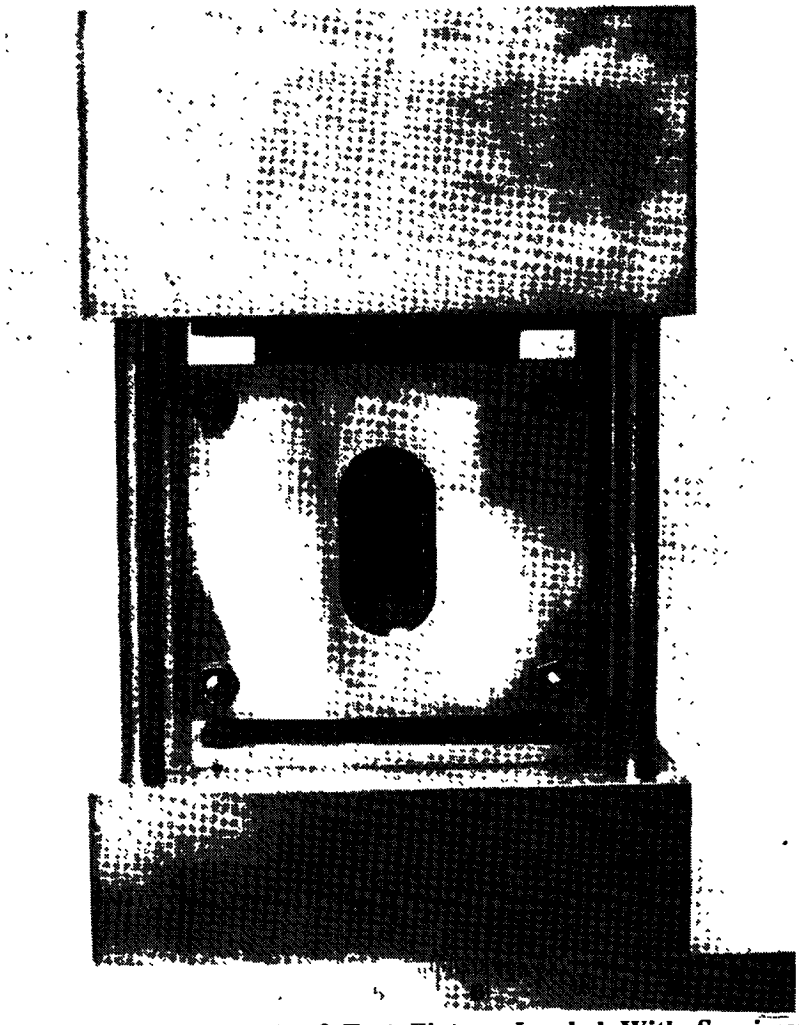


Figure 2. Photograph of Test Fixture Loaded With Specimen

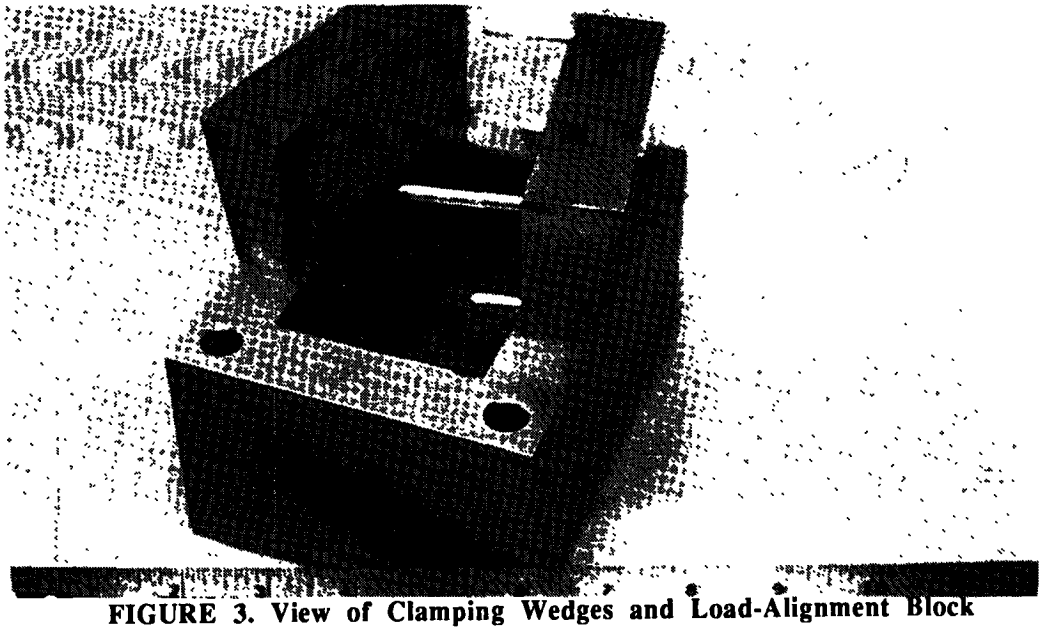


FIGURE 3. View of Clamping Wedges and Load-Alignment Block

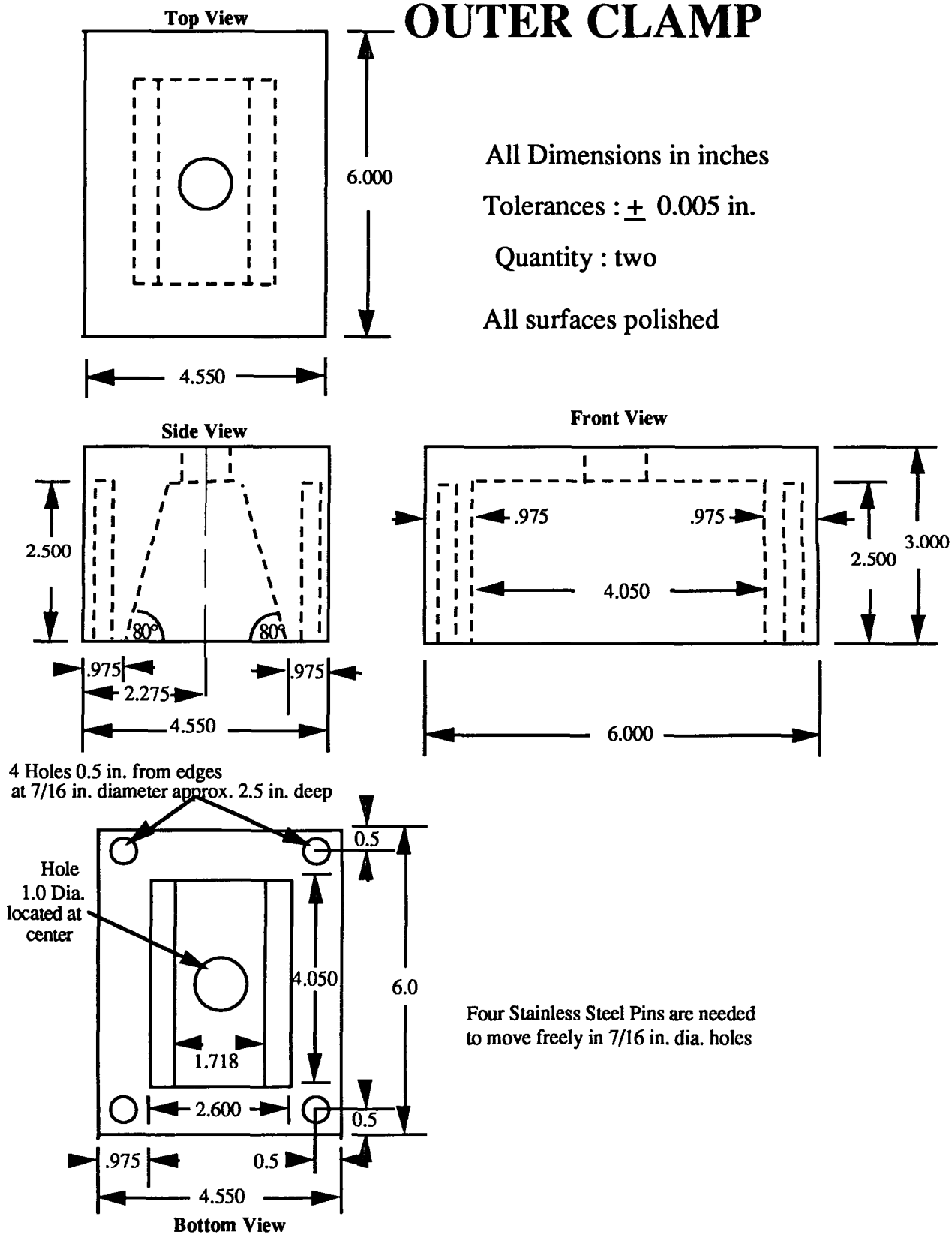


FIGURE 4. Detailed Drawing of Outer Clamp Piece

ANTI-BUCKLING FACEPLATES

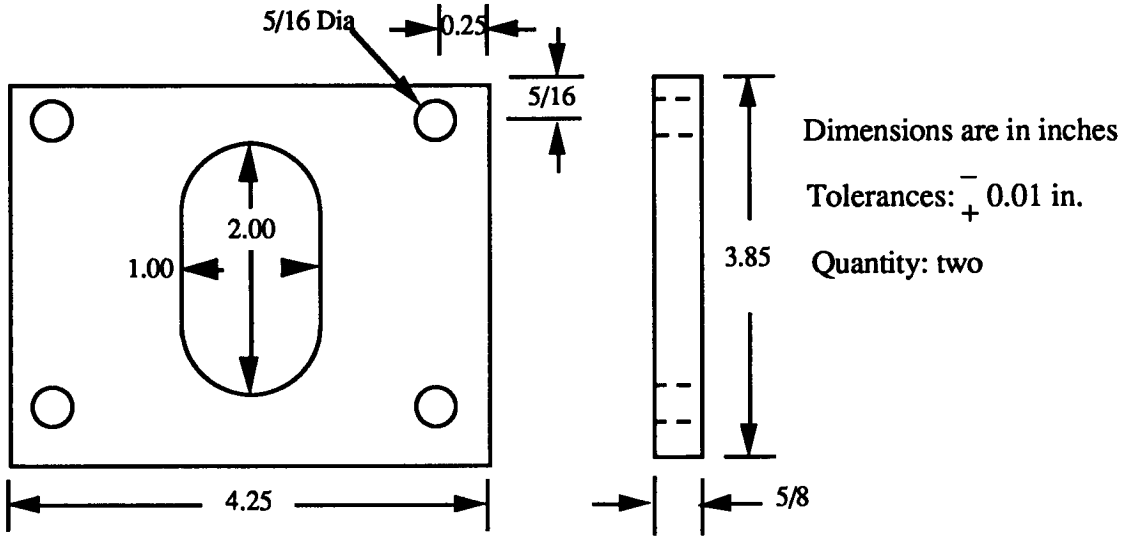
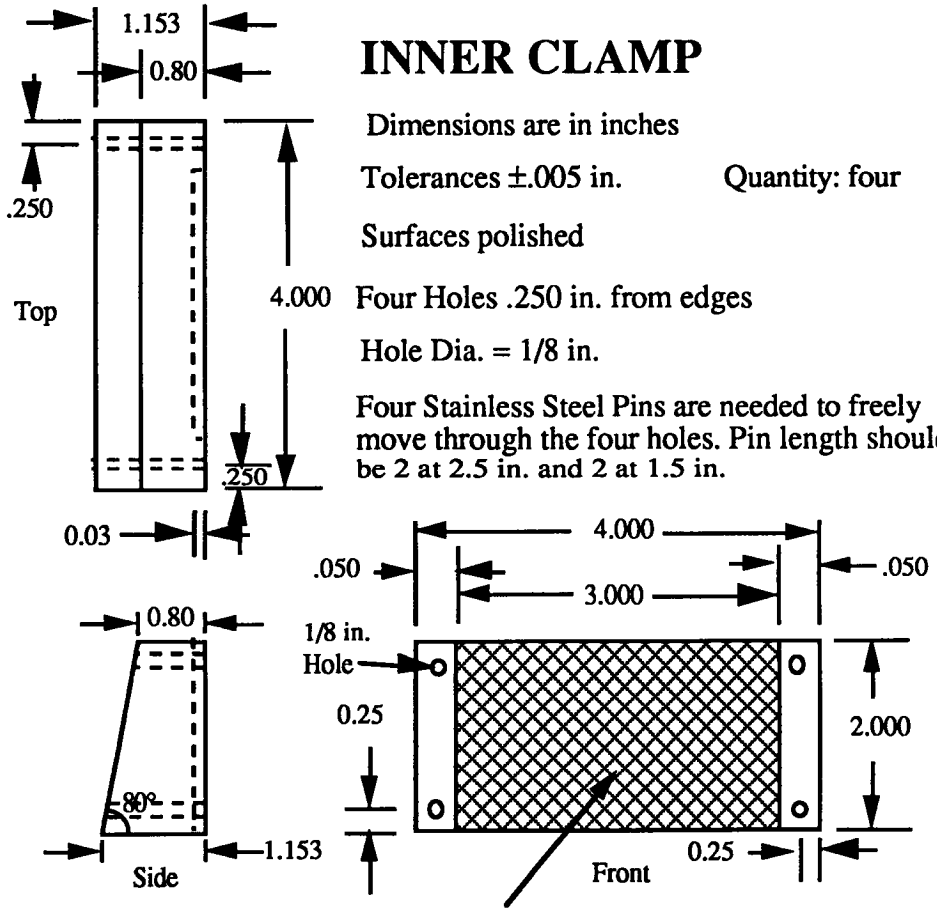


FIGURE 5. Detailed Drawing Of Anti-Buckling Faceplate.



INNER CLAMP

Dimensions are in inches

Tolerances ±.005 in. Quantity: four

Surfaces polished

Four Holes .250 in. from edges

Hole Dia. = 1/8 in.

Four Stainless Steel Pins are needed to freely move through the four holes. Pin length should be 2 at 2.5 in. and 2 at 1.5 in.

FIGURE 6. Detailed Drawing of Inner Clamp

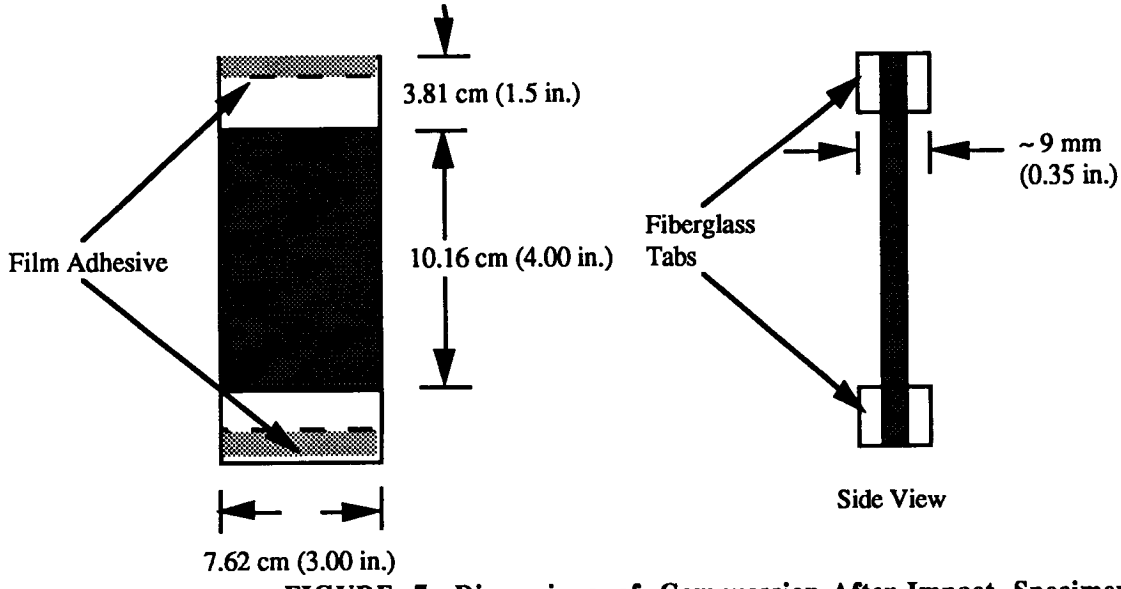


FIGURE 7. Dimensions of Compression-After-Impact Specimen

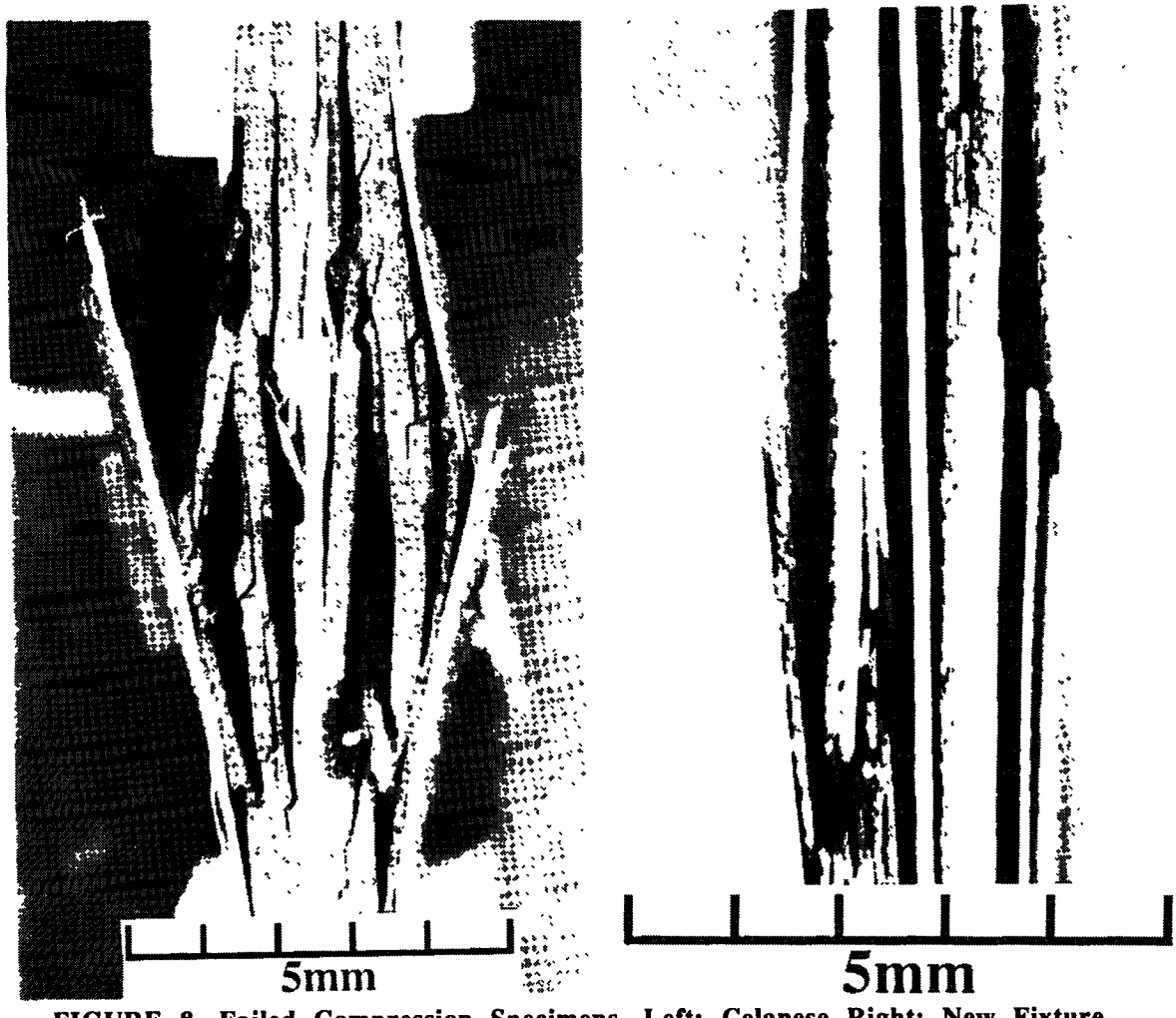


FIGURE 8. Failed Compression Specimens. Left: Celanese Right: New Fixture

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

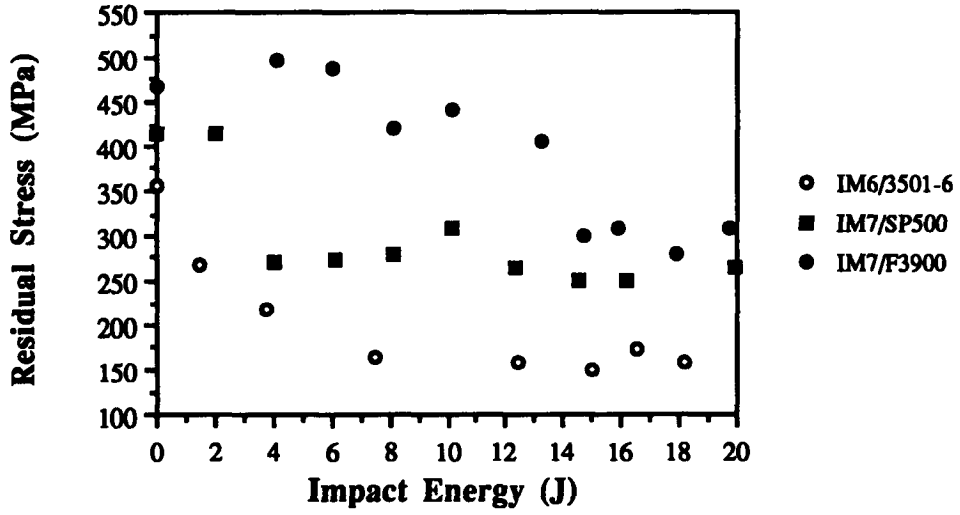


FIGURE 9. Residual Stress vs Impact Energy for Three Material Systems

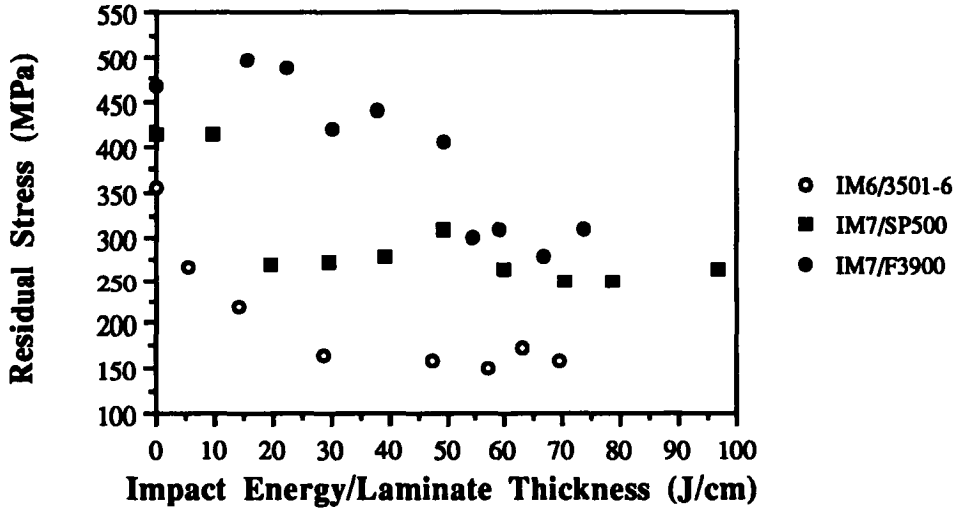


FIGURE 10. Residual Stress vs Impact Energy per Laminate Thickness

N92-22680

RESONANT ACOUSTIC DETERMINATION OF COMPLEX ELASTIC MODULI

David A. Brown and Steven L. Garrett, Department of Physics
Code PH/Gx, Naval Postgraduate School

ABSTRACT

A simple, inexpensive, yet accurate method for measuring the dynamic complex modulus of elasticity is described. Using a "free-free" bar selectively excited in three independent vibrational modes, the shear modulus is obtained by measuring the frequency of the torsional resonant mode and the Young's modulus is determined from measurement of either the longitudinal or flexural mode. The damping properties are obtained by measuring the quality factor (Q) for each mode. The Q is inversely proportional to the loss tangent (tan δ). The viscoelastic behavior of the sample can be obtained by tracking a particular resonant mode (and thus a particular modulus) using a phase-locked-loop (PLL) and by changing the temperature of the sample. The change in the damping properties is obtained by measuring the in-phase amplitude of the PLL which is proportional to the Q of the material. The real and imaginary parts or, in general, the complex modulus can be obtained continuously as a function of parameters such as temperature, pressure, or humidity. For homogeneous and isotropic samples only two independent moduli are needed in order to characterize the complete set of elastic constants, thus, values can be obtained for the dynamic Poisson's ratio, bulk modulus, Lamé constants, etc.

INTRODUCTION

The accurate measurement of the elastic constants of materials and their dependence on temperature, static pressure and other ambient parameters is important in many fields of science and engineering research as well as in product design and quality control. The elastic modulus relates the strains to the applied stresses, so its value and its temperature dependence are important design parameters. Manufacturer's specifications for elastic constants of castable polymers and epoxies usually are determined by static techniques, if at all, and rarely contain more than one modulus. The data also varies widely depending upon sample preparation and cure temperature, and never contain information about the temperature dependance of the moduli or their dynamic properties. Since it is the dynamic modulus which is needed for engineering applications involving noise, shock, vibration, transducer design, and acoustics, it is important that the dynamic properties are available to the designer.

Due to the number of variations in the preparation and large numbers of epoxies and polymers that are available from manufacturers, it is desirable to have a convenient technique for measuring the moduli of material samples. In the following sections a technique for measuring the dynamic properties of materials is described which is convenient, accurate, precise, and economical. The method relies on the measurement of the frequencies of the longitudinal, flexural, and torsional resonant modes of a single rod shaped sample using the same two transducers to excite and detect all three modes.

The fact that the technique is resonant insures high signal-to-noise ratio while the fundamental measurement being a frequency means that one can obtain extremely high precision with inexpensive instrumentation (i.e. a frequency counter). The technique can be used with both insulating or conducting samples which are not ferromagnetic. An additional attractive feature of this technique is the fact that in addition to the shear modulus, the Young's modulus can be measured independently using either or both the longitudinal and flexural modes.

The technique for measurement of the torsional mode is a refinement of one developed first by Barone and Giacomini [1] to study the modes of vibration of bars having variable cross-section. A similar arrangement was used later by Leonard [2] to disprove the existence of the "Fitzgerald Effect" [3-5] by measuring the attenuation of torsional waves in Teflon™. The technique was modified by Professor Isadore Rudnick at UCLA in order to excite and detect the flexural and longitudinal modes in addition to the torsional mode while still using the same transducer. He then incorporated the free-free bar technique as a teaching laboratory experiment in an upper level undergraduate acoustics class at UCLA. The technique has been extended to measure the complex modulus in a manner similar to that used by Barmatz, et. al.[6], by measuring the quality factor, Q, or free decay time. We have also included a phase-locked-loop which allows the continuous tracking of the moduli and loss tangent as a function of an external variable such as temperature [7-9].

MEASUREMENT TECHNIQUE

Modes of a Bar: A Rod of Circular Cross-Section

A uniform, bar-shaped sample of a homogeneous, isotropic solid having a circular cross-section of diameter, d, and length, L, will propagate three independent waves (torsional, flexural, and longitudinal) provided the wavelength of vibration, λ, is much greater than the bar diameter. The vibrational modes will exhibit characteristic resonances at particular frequencies which depend upon the dimensions of the sample, the density, and the elastic modulus. Of course, the resonant frequency will depend on the boundary conditions imposed on the ends of the rod. The simplest and most reproducible boundary condition to impose on the ends of the rod is that of zero stress and zero moment (i.e. a free-free boundary condition).

The torsional resonances of a bar having a free-free boundary condition are given by

$$f^T_n = \frac{n}{2L}\sqrt{\frac{G}{\rho}}\ ;\quad n = 1, 2, 3...\ , \tag{1}$$

where G is the shear modulus, ρ is the mass density, and L is the length of the rod. The dynamic shear modulus of the bar material can then be easily expressed in terms of its density, length, and the ratio of the frequency of its n-th mode of vibration to its mode number, n as follows:

$$G = 4\rho L^2\left(\frac{f^T_n}{n}\right)^2\ . \tag{2}$$

This result, in general, is dependent on the cross-sectional shape of the rod.

The longitudinal resonances of a bar having a free-free boundary condition are given by

$$f^L_n = \frac{n}{2L}\sqrt{\frac{E}{\rho}}\ ;\quad n = 1, 2, 3...\ , \tag{3}$$

where E is the Young's modulus of the rod material. This result, in general, is also independent of the cross-sectional shape of the rod as long as the initial assumptions ($\lambda >> d << L$, homogeneous, isotropic) are met. The dynamic Young's modulus of the bar can then be easily expressed in terms of its density, length, and the ratio of the frequency of the n-th mode of vibration to its mode number, n as follows:

$$E = 4\rho L^2\left(\frac{f^L_n}{n}\right)^2\ . \tag{4}$$

Unlike the torsional and longitudinal modes, the flexural waves of the bar obey a fourth-order differential equation and the flexural wave phase speed, c_F, is dispersive. The application of free-free boundary condition for flexural vibrations leads to a series of modes, f^F_n, given by

$$f^F_n = \frac{\pi B_n^2 c_F \kappa}{8L^2}\ ;\quad B_n = 3.0112, 4.9994, 7, 9, 11...., \tag{5}$$

This result is accurate at low frequencies where the effects of rotary inertia and shear deformations associated with the flexure can be neglected. The flexural wave phase speed, c_F, can be expressed as

$$c_F = \sqrt{2\pi f \kappa c_L}\ . \tag{6}$$

where κ, the radius of gyration, is given by

$$\kappa^2 = \left(\frac{1}{S}\right)\int z^2\, dS \tag{7}$$

Here S is the cross-sectional area of the rod and z is the distance of an element above the neutral axis in the direction of flexure. For a rod of circular cross-section $\kappa = d/4$. The Young's modulus of a bar in terms of its flexural resonances can be expressed as

$$E = \frac{1024}{\pi^2}\frac{\rho L^4}{d^2}\left(\frac{f^F_n}{B_n^2}\right)^2\ . \tag{8}$$

The redundancy provided by the fact that Young's modulus can be determined by the measurement of either the longitudinal or flexural modes provides a self-consistency check on the results and extends the frequency range over which the Young's modulus can be determined. This is a result of the flexural resonances occurring at frequencies which are typically an order-of-magnitude lower than the longitudinal resonances.

In using the ideal free-free boundary condition, the theoretical predictions do not take into account the added mass of the transducers which are placed on the bar to excite and detect the characteristic bar motion. However, the added mass is generally a small effect, typically a few percent of the entire mass of the rod. Also, the additional mass loading can easily be accounted for by the introduction of an effective length, L_{eff}, which provides a first-order correction if accuracy of greater than a few per cent is required. For the longitudinal and flexural modes, the additional mass loading of the transducers at the end of the bar has the same effect as lengthening the bar in the ratio of M : M + dM, where M is the mass of the "bare" rod and dM is the added mass of the transducer coils and their adhesive [10,7]. For the torsional mode, the effective correction is twice as large since the coils, mounted on the surface of the rod, make a proportionately greater contribution to the moment of inertia [7].

Selective Excitation and Detection of the Resonant Modes

The simplicity, speed, and economy of the free-free resonance bar method of modulus measurement is due to the ability to selectively and strongly excite the three resonant modes independently using the same inexpensive transducers. An electrodynamic transduction scheme is used to detect and excite all three modes using a pair of continuous coils of wire attached to each end of the sample bar. The sample is placed in various orientations within the field of a permanent magnet in order to selectively excite a particular mode. A typical apparatus for making these measurements is pictured in Figure 1, which shows the two magnets into which the coils are placed and an adjustable support structure in the foreground. The electronic interface (driver and signal amplifiers) is shown in the background.

The direction of the resulting differential Lorentz force, $d\vec{F}$, produced on a segment of wire, $\vec{dl}$, carrying a current, I, in a static magnetic field, $\vec{B}$, is given by

$$d\vec{F} = I\, \vec{dl} \times \vec{B}. \qquad (9)$$

It is the cleaver placement of the transducer wire coils, themselves attached to the bar, that allows for the excitation of one particular mode of vibration. Figure 2 illustrates the arrangement of the magnet and coil for excitation and transduction of the torsional modes. Here, the bar is placed with its axis at the center of the pole-piece faces which are aligned along the bar axis. The normal to the plane of coil is perpendicular to the magnetic field lines.

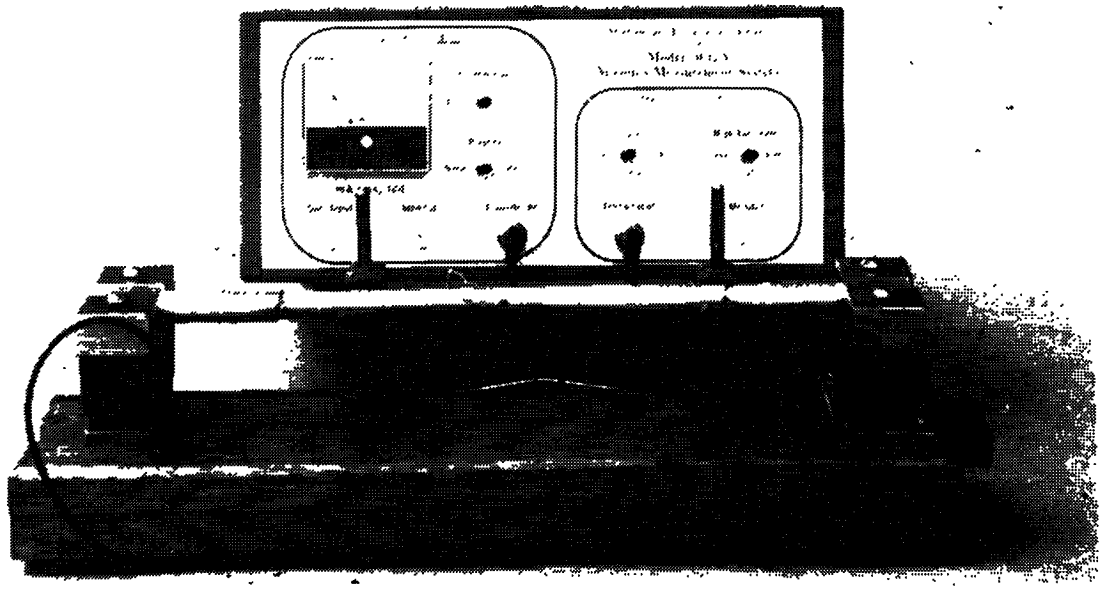
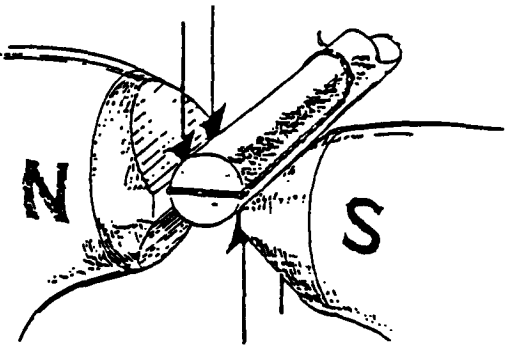


Figure 1. A measurement apparatus [11] used to electrodynamically excite and detect the flexural, longitudinal, and torsional modes of a free-free bar for determination of the shear and Young's elastic moduli.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

a)



b)

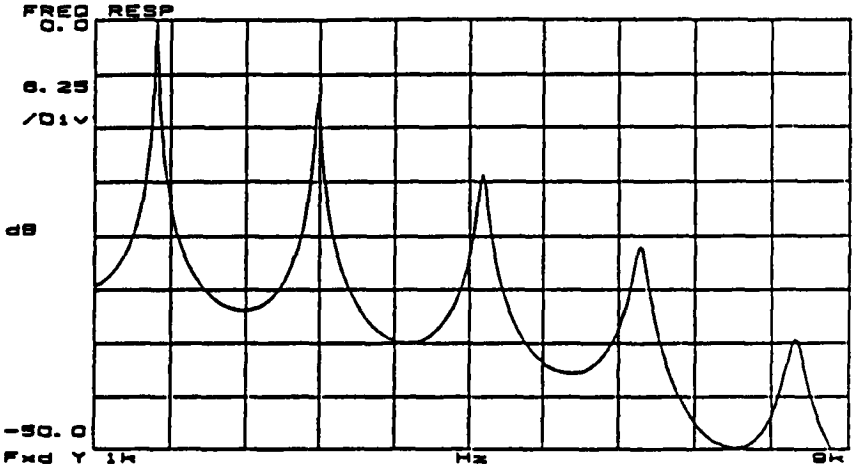


Figure 2. Torsional mode coil orientation and frequency response. (a) Orientation of the coils in the gap of the magnet pole pieces used to excite and detect the torsional mode. The arrows indicate the direction of the electromagnetic torque on the coil for a particular phase of the alternating electric current. (b) Frequency response output (log amplitude vs. frequency) for the five lowest torsional modes displayed on a HP 3562A Dual Channel Dynamic Signal Analyzer for an epoxy sample E-CAST™ F-28.

At the opposite end of the free-free bar, the receiver generates its "emf" as a result of the plane of the coil "rocking" back and forth in a region of strong, fairly uniform magnetic field. This change in angle modulates the flux through the coil since the projection of the area which the coil presents in the direction of the magnetic field is varying harmonically in time. Mathematically, we can express this induced voltage as

$$V = -\frac{d}{dt}\int_S \vec{B}\cdot\vec{n}\,dA \ , \tag{10}$$

where dA is in an incremental area subtended by the wire loop. For a small segment of wire moving with velocity, $\vec{u}$, in a magnetic field, $\vec{B}$, the induced emf is given by

$$emf = \vec{B}\cdot\vec{l}\times\vec{u} \tag{11}$$

The flexural mode can be observed by rotating the bar by 90° and translating it up or down by a distance approximately equal to the diameter of the bar. This places one of the long coil sections near the top of the rectangular pole faces while the other section is above (or below) the pole face and hence in a position of weaker magnetic field. The difference in the two opposing forces (the gradient) causes the bar to flex. Likewise, the receiver coil is then raised and lowered through the field gradient inducing a change in the flux through the coil and generating an emf.

For transduction of the longitudinal mode, the separation between the two magnet structures at either end of the bar is increased so that the strong region of magnetic field is concentrated primarily near the short section of coil which crosses the end of the bar along its diameter. The currents in the short section of the coil generate longitudinal forces on the end of the bar which excite the longitudinal resonance modes. Similarly, at the receiver, the coil is moved in and out of the strong field region by the wave-induced motion which generates the observed emf.

VISCOELASTICITY AND CONTINUOUS TRACKING OF THE STORAGE AND LOSS MODULUS

Dynamic (Storage) Modulus as a Function of Temperature

The dynamic modulus of elasticity of materials is important to engineers and scientists. These properties depend dramatically on temperature and it is their temperature dependance that often becomes critically important. The resonant bar technique described in this paper can be used for tracking both the storage and loss components of either the complex dynamic shear or Young's modulus. The tracking of the storage modulus is accomplished by tracking a particular resonant mode with a phase-locked-loop (PLL). The PLL is used to keep the measurement system "closed" and locked on resonance as the temperature of the sample is changed. If the modulus is a function of temperature, the resonance frequency of the bar will necessarily change. Figure 3 is a block diagram of the typical instrumentation used for such a automated tracking system.

After identifying the modes of the bar at room temperature a particular mode is selected for automatic tracking as a function of temperature. One adjusts the voltage controlled oscillator (VCO) manually to resonance with the error signal feedback path open so that there is no error signal presented to the voltage control (feedback) input. The phase shifter is then adjusted so that there is zero output from the quadrature signal channel of the lock-in amplifier at the preliminary setup temperature. As the temperature is then changed, a shift in the resonance frequency of the bar occurs, and the lock-in amplifier registers a quadrature output voltage that is used as a feedback signal to adjust the VCO. The quadrature signal changes the frequency output of the VCO to the same frequency and phase of the new resonance of the bar. It is this VCO output frequency that is proportional to the square of the appropriate modulus. The resonance frequency can be read from a suitable "bus compatible" VCO or frequency counter and the temperature of the samples with a small thermistor attached to a "bus compatible" multimeter so that the entire set of temperature and frequency measurement data can be acquired, displayed, and analyzed by computer. A program which controls the acquisition, analysis, and display of the data using an HP 9836 computer is included as an appendix to Reference [12] with modifications to track the loss modulus included in an appendix to Reference [9]. A sample data set for PR1592 is provided in Figure 4.

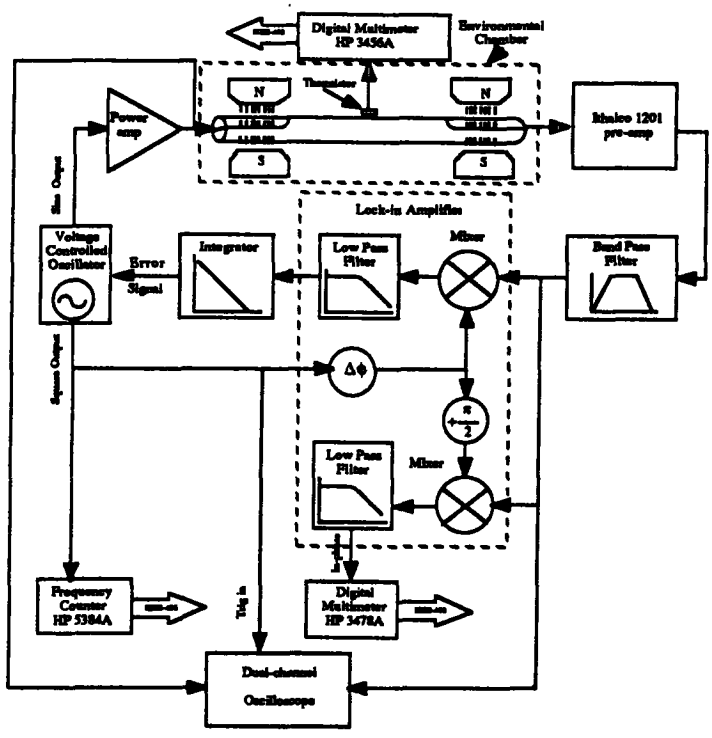


Figure 3. Block diagram of the phase-lock-loop used to automatically track the change in the resonance frequency (and thus modulus) of a particular mode as a function of temperature. The change in the loss tangent as a function of temperature can also be tracked by monitoring the in-phase output of the (lock-in) detector.

Dynamic Loss Tangent as a Function of Temperature

The loss tangent is proportional to the ratio of the energy dissipated in a material to the energy stored in the material in a given cycle of applied stress. The loss tangent is also equal to the inverse of the quality factor, Q, for a particular vibrational resonance. The temperature dependance of the loss tangent is obtained by measuring the in-phase voltage from the lock-in analyzer. This is possible since the response (in-phase voltage measured by the lock-in) of the vibrating rod at resonance is proportional to the Q of the system. Thus, by locking into a resonance as explained in the previous section by using the quadrature signal output of the PLL as a feedback signal to a VCO, the system stays on resonance independent of the temperature. The in-phase voltage is then monitored as a function of changing temperature to yield the temperature dependance of the Q. The inverse of the in-phase voltage is proportional to the loss tangent. The measurement of Q introduces additional experimental complications since care must be taken to insure that losses through the suspension system used to support the free-free bar are not significant compared to those intrinsic to the sample material under study. Because the strain distribution is known for each mode, these suspension losses can usually be reduced to insignificant levels since the two support points for the bar can be adjusted to occur arbitrarily close to velocity nodes.

Mathematically, the relationship between loss tangent and in-phase voltage can be shown through the derivation that follows which is based on a lumped oscillator model. The mechanical impedance of the bar at resonance is equal to the resistance of the system (by definition the reactance goes to zero at resonance) and can be expressed as

$$\frac{F}{u} = R , \tag{12}$$

where F is an applied force, R is the resistance, and u is the resulting velocity. The Q of the system can be expressed as

$$Q = \frac{\omega_o m}{R} , \tag{13}$$

where ω_o is the radian resonance frequency of the system, and m is the mass. Combining these two equations we can express the Q as

$$Q = \frac{u\omega_o m}{F} . \tag{14}$$

Now turning our attention back to the bar, suppose we drive the bar at resonance with a constant amplitude force, the ratio of the Q to the resonance frequency and velocity product remains constant, and equal to

$$\frac{Q}{u\omega_o} = \frac{m}{F} . \tag{15}$$

Recall that in the previous section the electrodynamic detection of the wire coil attached to the bar sample and suspended in the magnetic field, an emf was produced that is proportional to the velocity of the end of the bar. Letting V_{in} equal the in-phase component of the output voltage of the lock-in analyzer for a particular frequency and temperature, there is a constant of proportionality, A, that can be measured for the transduction scheme,

$$\frac{Q}{V_{in}\,\omega_o} = A . \tag{16}$$

Thus, as soon as this constant has been measured the loss tangent, tan δ, can then be determined as a function of the change in temperature and/or frequency of the sample,

$$\tan\delta = \frac{1}{Q} = \frac{1}{A_{in} V \omega_o} \tag{17}$$

The loss tangent can be measured directly at a particular frequency and temperature by any conventional method such as the 3dB down frequency technique or by measuring the relaxation time constant in a free decay mode. After having obtained the loss tangent directly the technique described above can be use to track the relative change in the modulus as a function of some external parameter variation such as temperature. A typical plot of the loss tangent is provided in Figure 4b. It is conventional [13] to plot the modulus and loss tangent as a function of reduced frequency $a_t f$ in order to obtain a master curve that illustrates the viscoelastic behavior of the material. Such a plot is provided in Figure 5.

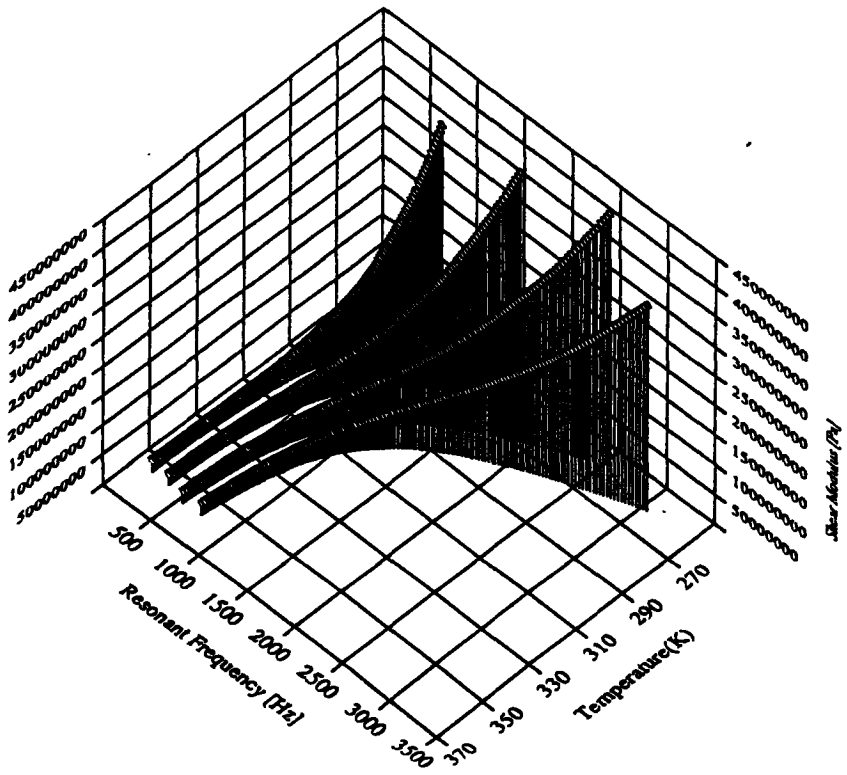


Figure 4a. Plot of the shear modulus of PR1592, as a function of temperature and frequency. Data was obtained for the shear modulus by tracking the first four torsional modes (separately) as the sample's temperature was changed. Only the first mode is shown for the loss tangent.

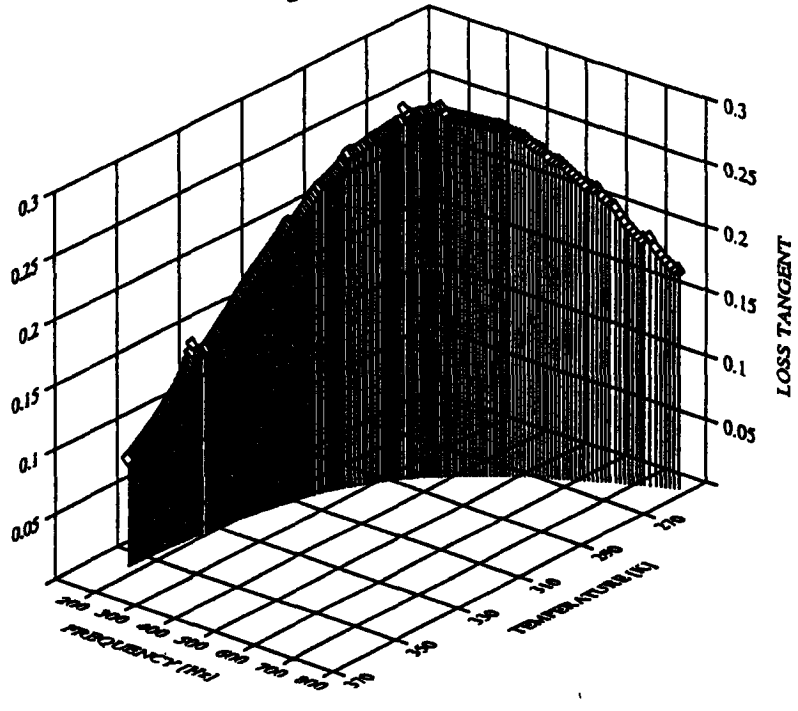


Figure 4b. Plot of the loss tangent of PR1592, as a function of temperature and frequency. Data was obtained as the sample's temperature was changed for the first torsional mode.

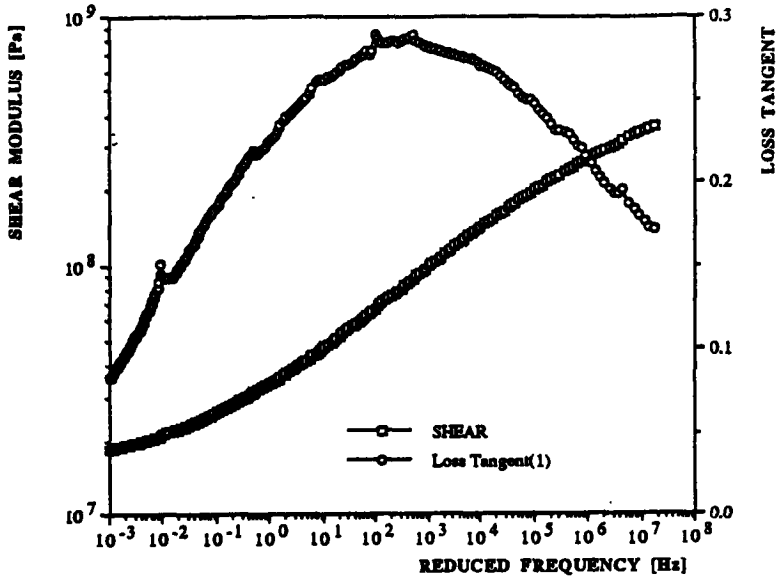


Figure 5. Master curve of loss tangent and storage shear modulus of PR1592 as a function of reduced frequency, $a_t f$, for a frequency shift parameter [13,14] log a_t = -12.9(T-283.15)/(107+T-283.15).

Comparison to the "Transfer Impedance" Technique

The advantages that this resonance bar technique has over the other conventional "transfer impedance" technique such as those discussed by Norris and Young [15] and Parsons, Yater, and Schloss [16] are briefly discussed here. The transfer impedance technique measures the amplitude and phase of the transfer function between typically two accelerometers mounted at either end of a vertically suspended rod-shaped sample which is driven from the upper end at fixed displacement by a shaker. One problem that this technique has is that the mass of the accelerometers is certainly not negligible as compared to the mass of the sample. What results is a difficult transcendental equation describing the motion of the bar. A computer is required to "invert" the frequency response information to account for this end-loading due to the accelerometer". In the Norris version, a Newton-Ralphson technique is used for this inversion. In both methods, transverse vibrations tend to be an undesirable problem and most importantly neither version of the transfer impedance method can measure the shear modulus. Thus, complete characterization of even the simplest materials, having only two independent moduli, is not possible. Other disadvantages include the fact that the sample must be clamped or otherwise bonded to the shaker at one end and the accelerometer must also be attached at the other end. Also, the shaker table and accelerometer is orders of magnitude more costly than the electrodynamic coil-magnet technique described in this paper.

Another important issue is reproducibility. The shaker table and accelerometer attachments can introduce gross ireproducibilities and misleading results due to the bonding and clamping of these fixtures. Since the cost of the accelerometers is not small, they are typically removed after each measurement. And, for accurate remeasurement one requires that the bonding be duplicated exactly. If one attempted to get a second independent moduli such as the shear modulus one would require, in addition to very accurate measurements, that the second (entirely different) apparatus would produce the same compensating error in order to obtain other elastic constants such as Poisson's ratio.

The transducers cost pennies and weigh less a penny (typically a few grams each, including the adhesive). Because the transducer is so cheap, there is no reason to remove them from samples hence re-testing, even years later, can resolve small or slow changes which would be masked by the remounting more massive accelerometers. Both the Young's and shear moduli can be measured using the same apparatus and the same pair of transducers. The ability to excite and detect both the longitudinal and flexural modes provides a redundancy in the determination on the Young's modulus that checks the method for self-consistency and/or extends the frequency range over which the measurement is possible. In addition to the storage modulus, the loss tangent is also easily accessible with this resonant bar technique.

ACKNOWLEDGEMENTS

This work was funded in part by the Naval Postgraduate School Direct Funded Research Program and the Naval Sea Systems Command. The data on PR1592 was obtained by Beng-Hock Tan, a graduate student from the Sinapore Department of Defense.

REFERENCES

1. A. Barone and A. Giacomini, "Experiments on some electrodynamic ultrasonic vibrators", Acoustica **4**, 182-184 (1954).
2. R. W. Leonard, "Attenuation of torsional waves in teflon", J. Acoust. Soc. Am. **40**(1), 160-162 (1966).
3. E. R. Fitzgerald, "Simple method for observing audiofrequency resonances and sound beams in crystals", J. Acoust. Soc. Am. **36**(11), 2086-2089 (1964), and references therein.
4. R. W. Leonard, "Comment on the existence of the Fitzgerald Effect", J. Acoust. Soc. Am. **38**(4), 1-2 (1965).
5. I. Rudnick, "Acoustics goofs or irreproducible effects in acoustics", J. Acoust. Soc. Am. **83**(Suppl. 1), S38 (1988).
6. M. Barmatz, H. J. Learny, and H. S. Chen, "A method for the determination of Young's modulus and internal friction in metallic glasses", Rev. Sci. Inst. **42**, 885-887 (1971).
7. S. L. Garrett "Resonant acoustic determination of elastic moduli", J. Acous. Soc. Am., **88** (1), 210-221 (1990).
8. D. A. Brown, Beng-Hock Tan, and S. L. Garrett, "Nondestructive dynamic complex moduli measurements using a Michelson fiber interferometer and a resonant bar technique" in Fiber Optic Smart Structures and Skins III, Proc. Soc. Photo-optical Inst. Eng. (SPIE), **1370**, pp. 238-247, (1990).
9. Beng-Hock Tan, "Resonant acoustic determination of complex elastic moduli", Master's thesis in Engineering Science, Naval Postgraduate School, Monterey, CA, March, 1991.
10. J. W. Strutt (Lord Rayleigh), "Some general theorems relating to vibrations." Proc. Math Soc. (London) **4**, 357-368 (1873); *Collected Works* (Dover, New York,1964), Vol. I, 170-181.
11. T. Lentz, "Year's best high-tech products earn R&D 100 recognition", Res. & Dev. **32**(10), 79-80 (1990);Mechanical Measurement Devices, Model 3012A Dynamic Modulus Measurement System, P.O. Box 8716n, Monterey, CA 93943.
12. K. Wetterskog, B. L. Beaton, and J. Serocki, "A fiber-optic acceleration canceling hydrophone made of a castable epoxy", Master's thesis in Engineering Acoustics, Naval Postgraduate School, Monterey, CA, June, 1990.
13. J. D. Ferry, *Viscoelastic Properties of Polymers*, 2nd ed., John Wiley & Sons, Inc., 1970.
14. R. N. Capps, *Elastomeric Materials for Acoustical Applications*, (Naval Research Laboratory - Underwater Sound Reference Detachment, 1989), p. 31-32.
15. D. M. Norris, Jr. and W. C. Young, Exp. Mech. **10**, 93 (1970).
16. J. S. Parsons, W. Yater, and F. Schloss, "The measurement of dynamic properties of materials using a transfer impedance technique", Naval Ship Research and Development Center - Acoustics and Vibration Laboratory Research and Development Report No. 2981 (April, 1969), DTIC Report No. AD 688786.

ROBOTICS

(Session C5/Room B4)

Wednesday December 4, 1991

- **A Unique Cable Robot for Space and Earth**
- **A Lightweight, High-Strength Dexterous Manipulator Arm**
- **Real-Time, Interactive Simulator System for Telepresence**
- **A Hazard Control System for Robot Manipulations**

N92-22681

THE CLIMBING CRAWLING ROBOT
(A UNIQUE CABLE ROBOT FOR SPACE AND EARTH)

James J. Kerley/754.1
Edward May/NSI
Wayne Eklund/NSI
Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

Some of the greatest concerns in robotic designs have been the high center of gravity of the robot, the irregular or flat surface that the robot has to work on, the weight of the robot that has to handle heavy weights or use heavy forces and the ability of the robot to climb straight up in the air. This climbing crawling robot handles these problems well with magnets, suction cups or actuators. The cables give body to the robot and it performs very similar to a caterpillar. The computer program is simple and inexpensive as is the robot. One of the important features of this system is that the robots can work in pairs or triplets to handle jobs that would be extremely difficult for single robots. The light weight of the robot allows it to handle quite heavy weights. The number of feet give the robot many roots where a simple set of feet would give it trouble.

INTRODUCTION

Through the years it has been difficult for robots to maneuver because the robot has a high center of gravity and the precision of its accuracy depends upon the angular rotation and the necessity of the grippers to precisely mate with and grab a target. This concept was developed first as a method for climbing, stripping and painting towers that were many feet off the ground.

An individual element is shown in Figure 1. It consists in two "U" structures held together by two circular cables. It is mounted on a magnet shown at the far left. The mounts for individual elements could be magnets, suction cups, grippers or wheels. The electronics are not shown but it will be discussed later.

FORWARD MOTION

The individual elements described in the introduction are attached by circular elements with a hole down the center to carry the air and the DC current. Eventually these elements will index left and right so that the robot can reach any structure in space with the proper orientation to move on.

The forward motion is shown in Figure 2. Step 1 shows the actuators in neutral while the magnets are all firmly attached to the structure. In step 2 the Number 1 magnet is released while its actuators are extended. Then the number 1 magnet is firmly attached. In step 3 the No. 1 and number 3 magnets are held while the number 2 actuators are extended. The number 1 actuators are limber. The number 3 actuators are held. In Step 4 the no.1 and no. 2 magnets are held while the no. 2 actuators pull no 3 along. This is just one sequence of motions. There are many more.

Figure 3 shows the robot going over an obstacle. This maneuver is accomplished by taking the first element and lifting it off the ground while the no. 2 element is held firm. Then the no 3 element lifts the no. 1 and no. 2 elements off the ground. This is continued until the first element goes over the obstacle. When the no. 3 and no. 4 elements go forward the number 1 elements has their actuators go into neutral. As it goes further the no. 1 element bends the other way until it strikes the structure. Then the forward part of the robot lifts the back parts over the obstacle.

PRECEDING PAGE BLANK NOT FILMED

There are many other maneuvers not discussed here. such as a climbing turn and lift. These get more complex. The feature that makes all of these maneuvers perform so well is that all of the robot is not standing on two or three grippers to perform a function.

CABLE ROBOT CONTROL ARCHITECTURE

This prototypes a standard off the shelf controlled which is generally used for the automated machine industry. The controller allows up to 16 outputs (on/off) and 8 inputs. A personal computer acts as a 'dumb' terminal, communicating to the interface controller in Figure 4. The programmer sends programs to the controller from the personal computer via serial links. The interface controller then executes the programs by simply stepping through the instructions. The instructions consist of 'states' that the outputs can be in during each step of the program, and its duration. Since the prototype will have a given number of known states, it is a simple matter to organize these states into coherent programs which will allow the Cable Robot to walk forward, grasping with the electromagnets at the appropriate times, and extending or retracting the segments in turn.

As example of a program to move the forward two magnets forward is as follows: (1 = actuator extended or magnet on)

STATE:	1a	1b	2a	2b	3a	3b	4a	4b	4c
Module 1:									
Act 1	0	1	1	1	0	0	0	0	
Act 2	0	1	1	1	0	0	0	0	
Act 3	0	1	1	1	0	0	0	0	
Act 4	0	1	1	1	0	0	0	0	
Magnet 1	0	0	1	1	1	1	1	1	
Module 2:									
Act 5	0	0	0	0	1	1	1	0	
Act 6	0	0	0	0	1	1	1	0	
Act 7	0	0	0	0	1	1	1	0	
Act 8	0	0	0	0	1	1	1	0	
Magnet 2	1	1	1	0	0	1	1	1	
Module 3:									
Act 9	0	0	0	0	0	0	0	0	
Act 10	0	0	0	0	0	0	0	0	
Act 11	0	0	0	0	0	0	0	0	
Act 12	0	0	0	0	0	0	0	0	
Magnet 3	1	1	1	1	1	1	0	0	

Note: This would be a "2/3" step forward, since the actuators in the third module never extend. Other sequences would allow larger steps to be taken, but for demonstration purposes, this will serve.

The prototype controller will allow the states to be changed as often as ten times a second, so that the above steps would take about a second. If we allowed the electromagnets to 'trade' in one program step, then steps 2b and 4a could be eliminated. This would allow the prototype to walk about 5 feet per minute.

Additional prototpype experiments would include inclined walk programs and turning walk programs in which two of the four actuators in a module would be extended while the other two are retracted. This allows an angle to be developed in any of four planes in the module: up, down, right, or left. The actual angle developed depends on the stroke of the actuator and the construction ratios of the module cable joint.

Connected to the computer outputs are circuits which convert the digital logic signals to coil drivers which actuate the miniature air valves for the cylinders. The same coil drivers will drive relays to turn on or off the electromagnets. The digital side is protected by opto-isolators, and the relays are protected by snubber circuits.

The computer program in the prototype will have several demonstration modes:

* Walk forward
* Walk backward
* Make right turn
* Make left turn

The modules can be outfitted with any one of a number of capabilities. Just a few are listed:

* Gripper actuation (as alternative to electromagnets)
* Vacuum cup (As alternative to electromagnets)
* Camera control (pan, tilt, focus, etc.)
* Sand blast
* Paint chipper
* Air blast cleaner
* Brush clean
* Vacuum cleaner head
* Paint (spray, roller, brush applicators)
* Sensor mount (metal detector, Distance measurement, etc.)
* Energy storage (battery, air cylinder)
* Radio communications
* Cable payout/retrieval

The modules can be made to operate under water with appropriate sealing of the electronics and using water or hydraulic actuation fluids. Bellows type covers can be fitted to protect from paint spray or sand blast damage.

Applications for this robot include:

* High tension wire tower (inspect, dye penetrant, weld, paint and repair)
* Drill rig inspection and repair
* Building air duct inspection/cleaning
* Overhead crane inspection
* Piping inspection (dry or wet)
* Hazardous waste container inspection and clean out
* Surveillance camera positioning
* Bridge girder inspection, repair and paint
* Radio tower inspection, repair and paint
* Ship's tanks inspection, repair and paint
* Boiler inspection, repair and paint or clean.

Connected to the computer outputs are circuits which convert the digital logic signals to coil drivers which actuate the miniature air valves for the cylinders. The same coil drivers will drive relays to turn on or off the electromagnets. The digital side is protected by opto-isolators, and the relays are protected by snubber circuits.

The computer program in the prototype will have several demonstration modes:

* Walk forward
* Walk backward
* Make right turn
* Make left turn

The modules can be outfitted with any one of a number of capabilities. Just a few are listed:

* Gripper actuation (as alternative to electromagnets)
* Vacuum cup (As alternative to electromagnets)
* Camera control (pan, tilt, focus, etc.)
* Sand blast
* Paint chipper
* Air blast cleaner
* Brush clean
* Vacuum cleaner head
* Paint (spray, roller, brush applicators)
* Sensor mount (metal detector, Distance measurement, etc.)
* Energy storage (battery, air cylinder)
* Radio communications
* Cable payout/retrieval

The modules can be made to operate under water with appropriate sealing of the electronics and using water or hydraulic actuation fluids. Bellows type covers can be fitted to protect from paint spray or sand blast damage.

Applications for this robot include:

* High tension wire tower (inspect, dye penetrant, weld, paint and repair)
* Drill rig inspection and repair
* Building air duct inspection/cleaning
* Overhead crane inspection
* Piping inspection (dry or wet)
* Hazardous waste container inspection and clean out
* Surveillance camera positioning
* Bridge girder inspection, repair and paint
* Radio tower inspection, repair and paint
* Ship's tanks inspection, repair and paint
* Boiler inspection, repair and paint or clean.

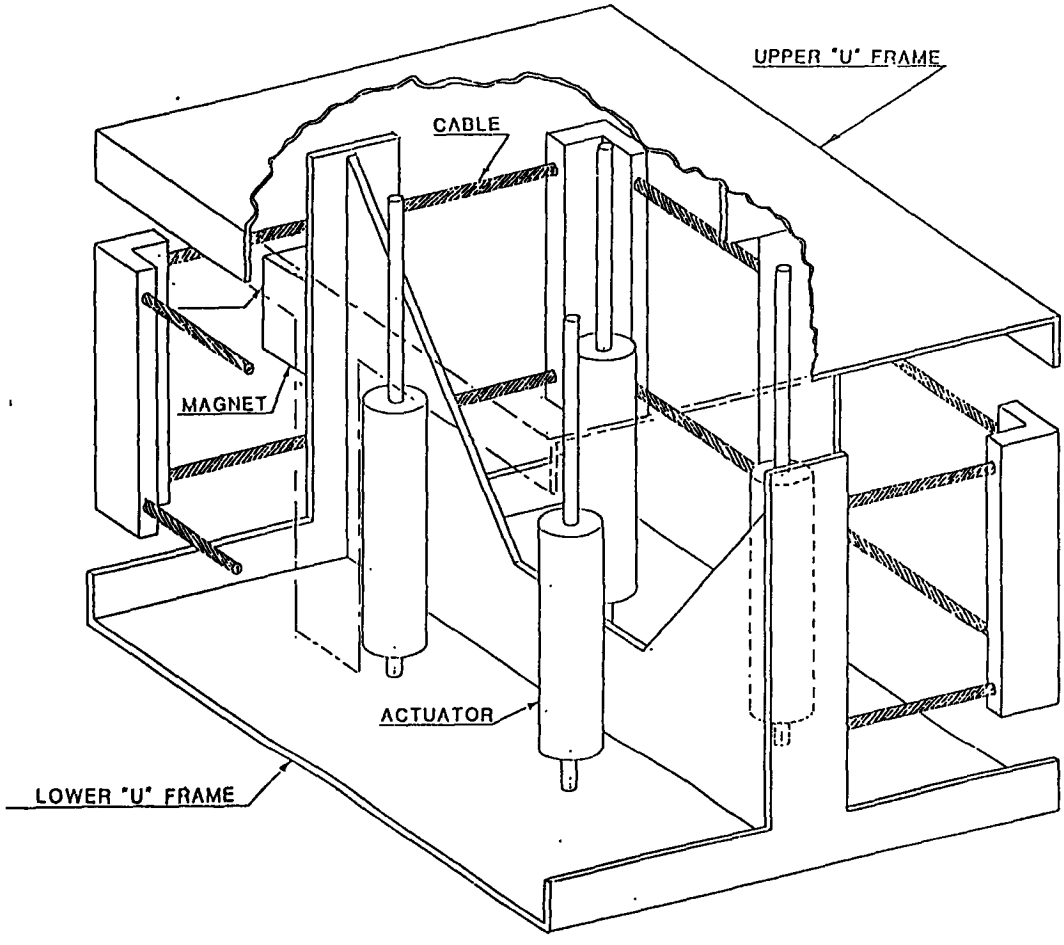


FIGURE 1 PERSPECTIVE

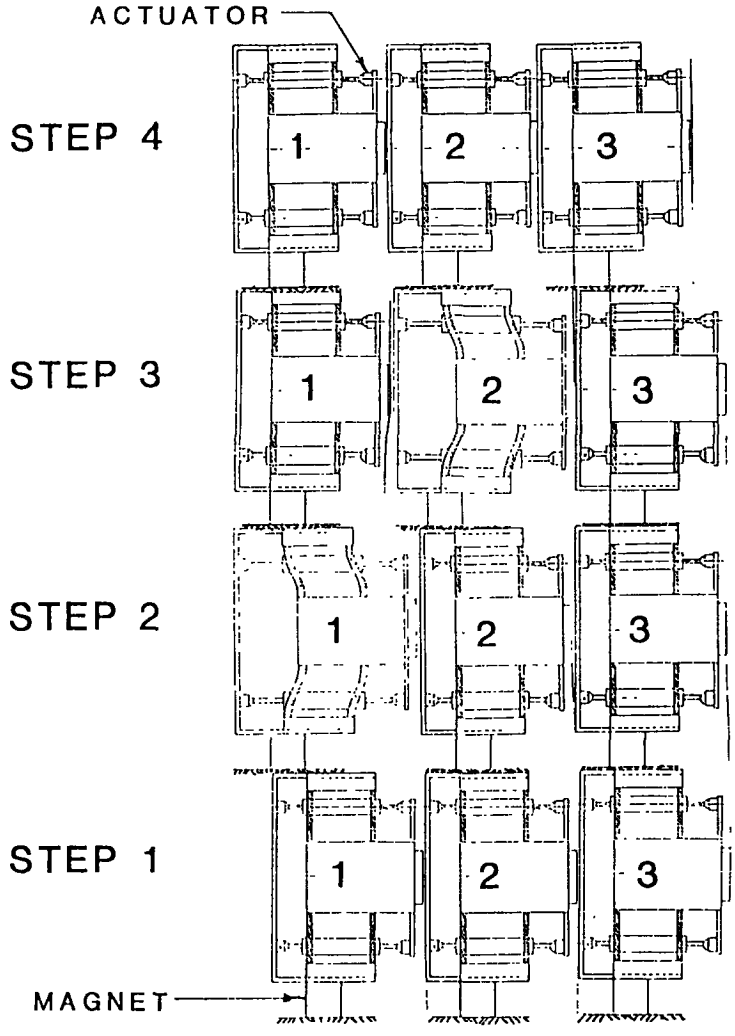


FIGURE 2 MOTION

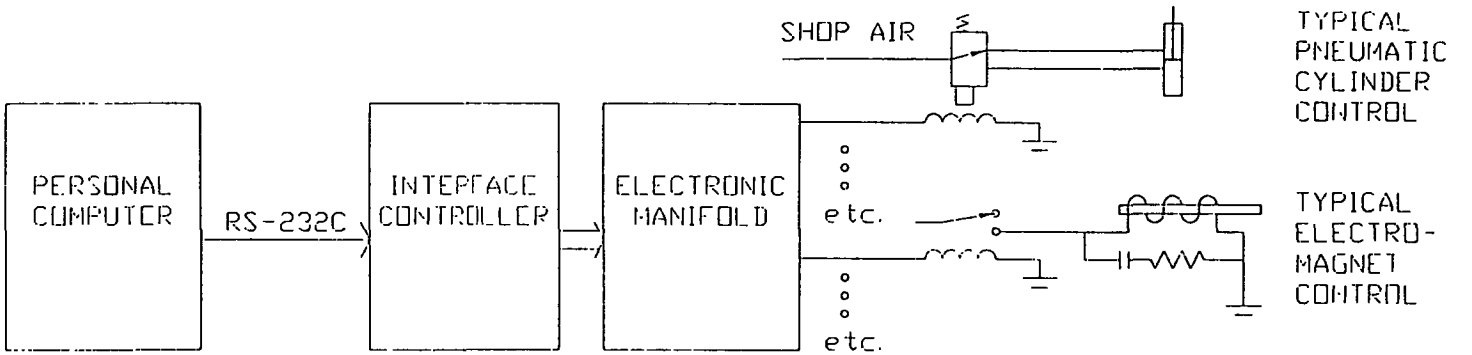


FIGURE 4 ELECTRONICS

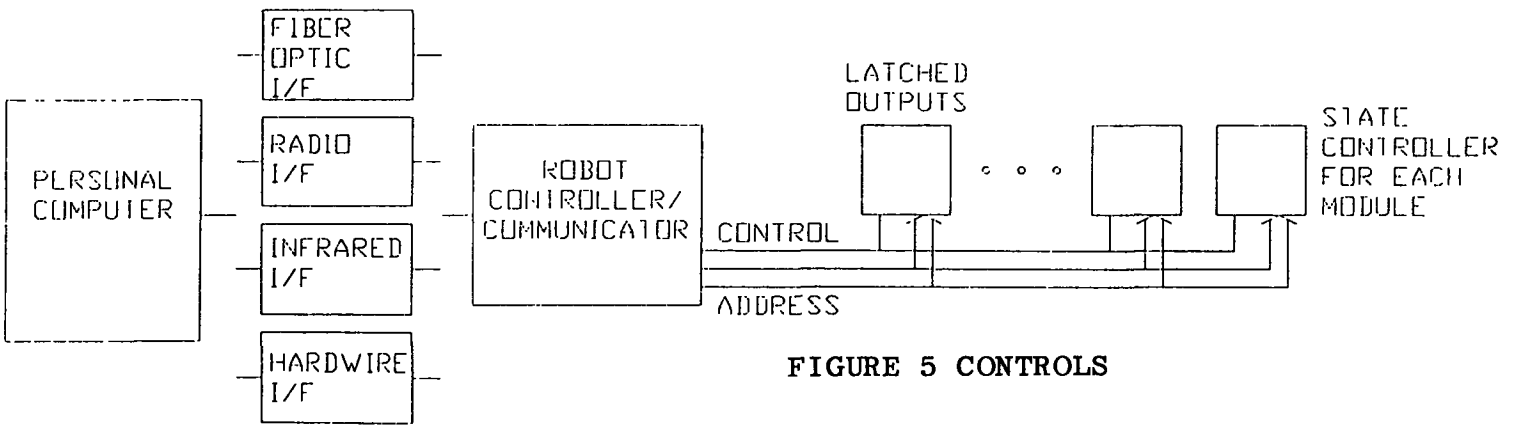


FIGURE 5 CONTROLS

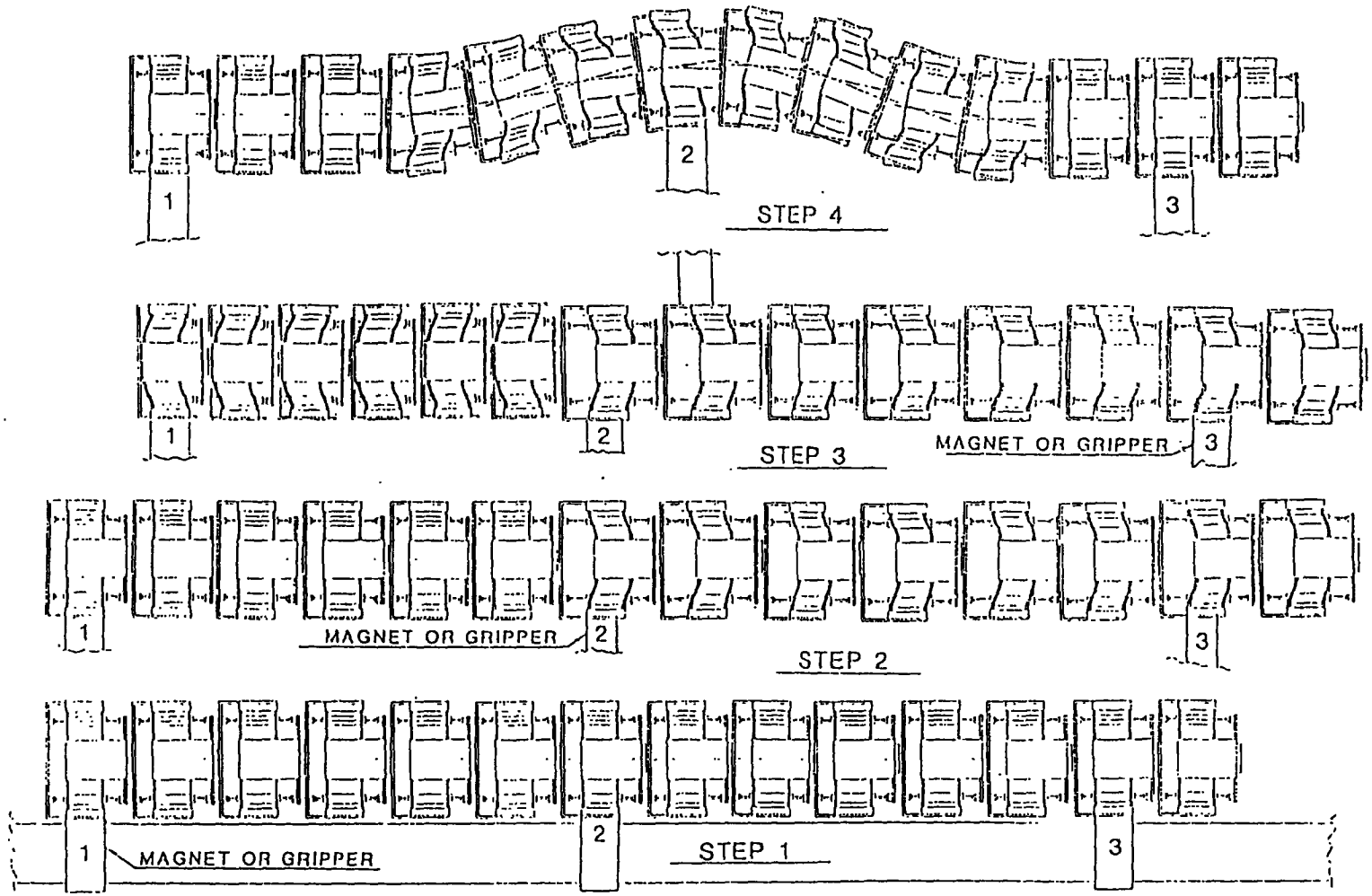


FIGURE 3 OVER AN OBSTACLE

N92-22682

A LIGHTWEIGHT, HIGH STRENGTH DEXTEROUS MANIPULATOR FOR COMMERCIAL APPLICATIONS[1]

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

Bruce M. Schena
Steve M. Cohan
Odetics, Inc.
1515 S. Manchester Avenue
Anaheim, CA 92802

ABSTRACT

This paper describes the concept, design, and features of a lightweight, high strength, modular robot manipulator being developed for space and commercial applications. The manipulator has seven fully active degrees of freedom and is fully operational in 1 G. Each of the seven joints incorporates a unique drivetrain design which provides zero backlash operation, is insensitive to wear, and is single fault-tolerant to motor or servo amplifier failure. Feedback sensors provide position, velocity, torque, and motor winding temperature information at each joint. This sensing system is also designed to be single fault-tolerant. The manipulator consists of five modules (not including gripper). These modules join via simple quick-disconnect couplings and self-mating connectors which allow rapid assembly/ disassembly for reconfiguration, transport, or servicing. The manipulator is a completely enclosed assembly, with no exposed components or wires. Although the initial prototype will not be space qualified, the design is well-suited to meeting space qualification requirements. The control system provides dexterous motion by controlling the endpoint location and arm pose simultaneously. There is access to the control system at multiple functional task levels. Potential applications are discussed.

INTRODUCTION

Odetics Inc. is developing a new, high performance manipulator that will address new market opportunities space, defense, and commercial applications. Although these applications are embryonic and ill-defined, current manipulators clearly lack the general performance capabilities these tasks will require. Recent research in space telerobotics has made dexterity, fault tolerance, and safety requirements clearer [1-3]. The general approach guiding this design is to build an advanced manipulator which uses the best ideas from existing designs and has new features required for advanced applications in both the space and commercial arenas. Sophisticated software and control algorithms have been developed concurrently with the manipulator hardware, yielding an integrated system that is adaptable to many applications.

This applications class excludes most conventional industrial robots. In fact, there exists only a few commercially available manipulators that are potentially suitable for applications in dynamic, unstructured environments such as space and hazardous material handling. Some of these machines are hydraulically powered, such as the manipulators

1. This work was partially funded by NASA Small Business Innovation Research Contract NAS-1062

produced by Schilling Development[1], Kraft Telerobotics[2], and Sarcos [4]. While hydraulic manipulators have very good strength, speed, and size characteristics, they require hydraulic power support hardware with its inherent size and weight penalty. There is the additional danger of flammable hydraulic fluid leakage. These detriments make hydraulic manipulators unsuitable for some applications, such as those in space. Perhaps the only commercially available electric manipulators similar to the Dexterous Manipulator are those produced by Robotics Research, Inc. [5]. Some of the differences between the two manipulators are described in this paper.

DESIGN OBJECTIVES

New market applications will require autonomous and teleoperated manipulation in unstructured, dynamic environments. The capabilities of the manipulator system will ultimately determine the success or failure of these operations. As with most system developments, cost and development time requirements must balance performance and reliability goals. Since definitions of the tasks to be performed are still evolving, a reconfigurable system that could be easily adapted to various applications would be attractive. In particular, commercial applications could benefit from the reduced cost of reconfiguring the system for a new application, in contrast with developing special equipment, such as tooling, for each new use. The system should be configured to fit the job, not vice versa.

These considerations led to the adoption of a modular manipulator architecture. A set of self-contained manipulator modules with standard interfaces provides lower cost and minimizes development time of specialized systems. In addition, modularity allows easy transportation to a remote location, fast on-site assembly, and quick in-the-field repairs. Useful configurations are not limited to manipulators - self-contained actuator modules can be configured into other application-specific mechanisms with fewer or more degrees of freedom.

Some specific mechanical design challenges arising from the modular architecture approach include:

- Mechanical and electrical module interfaces
- Component packaging and wire harness design
- Scalable actuator topologies.

More general mechanical design and engineering goals include:

- Maximum payload to weight ratio and compact design
- High dexterity
- Fault tolerant sensing and actuation
- Fully enclosed mechanisms and wiring
- Accurate joint torque sensing.

Design issues specific to the control of a high performance kinematically redundant manipulator include:

- Providing sensing for advanced control techniques
- Redundancy management, including singularity avoidance and configuration (pose) control
- Robustness and fault tolerance.

1. Schilling Development, Davis, CA
2. Kraft Telerobotics Inc., Overland Park, KS

Table 1 summarizes the principal performance goals.

Table 1 Manipulator Performance Goals

Attribute	Target Value	Notes
Length	55 in.	shoulder pitch to toolplate
Weight (1 G)	165 lb.	actual weight - 150 lb.
Max Endpoint Speed	> 40 in./s	for task space moves
Payload Lateral Force	50 lb. 20 lb. 135 lb.	peak - short duration continuous duty at toolplate, fully extended
Dexterity	7 active degrees of freedom	
Repeatability	0.025 in.	
End Effector Support	72 wires	72 to forearm; 40 to toolplate

Another important design objective was to create a system that could operate terrestrially as well as in a microgravity environment. Previous space manipulators were not operational in 1 G and required special equipment for ground testing. Within the financial scope of this effort, the immediate objective was to develop a system that is a reasonable design evolution away from becoming a space-qualified machine.

MANIPULATOR CONFIGURATION

Figure 1 shows the Dexterous Manipulator. This kinematic arrangement of joint modules includes two shoulder modules (azimuth and elevation), an upper arm roll module, an elbow module, and a three joint wrist module. The upper arm roll module allows the plane formed by the upper arm and forearm to rotate, providing capability for manipulator configuration control. Each joint has a large range of motion, providing a large, dexterous workspace. The elbow (joint 4) offset allows the lower arm to fold up against the upper arm, providing excellent manipulator stowage.

JOINT MODULES

Many of the innovative and unique features of the Dexterous Manipulator are apparent in the joint module design. Each module contains motors, sensors, wiring, transmission elements, and structure. Each joint uses exactly the same drivetrain concept, scaled according to that joint's torque requirements. Module interfaces consist of both positive mechanical connection and self-mating electrical connectors held together with simple clamping collars. There are no inter-module electrical connections that the user must make. This quick disconnect design allows the manipulator to be assembled or disassembled in approximately seven minutes. Table 2 shows the pertinent characteristics of each of the module types.

ACTUATORS AND TRANSMISSION

All of the joints use brushless D.C. motors for actuation. The actuator transmissions use spur gear technology with special mesh geometries and materials to obtain high torque capability. The unique drivetrain design uses a parallel topology that eliminates all backlash without gear mesh adjustments. Figure 2 illustrates the basic concept. In each joint, two motors drive a common output member. Under normal operation, one motor acts as the "prime mover",

Table 2 Module Performance Characteristics

Module	Range of Motion (deg)	Weight (lb)	Peak Torque (in-lb)	Peak Speed (deg/s)	Position Resolution (deg)
Shoulder Azim[a]	325	34.5	8000	72	6.25E-4
Shoulder Elev	325	34.5	8000	72	6.25E-4
Upper Arm Roll	717	27.5	4000	91	6.99E-4
Elbow	235	24.5	4000	91	6.99E-4
Wrist Pitch	238	1	1300	150	7.56E-4
Yaw	208	27.5	1300	150	7.56E-4
Roll	340	1	1300	150	5.50E-3

a. The two shoulder modules are identical.

providing the driving torque, while the second motor provides a small bias torque in the opposite sense. The bias torque removes all backlash from the transmission. Backlash remains zero through continued operation and wear, with no special adjustments required. When large torque s are required, the biasing motor reverses and provides additional torque, at the expense of zero backlash operation. The design also provides single fault tolerance to motor or motor driver failures. A joint can continue to operate in a controlled manner (with reduced bandwidth) after such a failure. After the task at hand is completed, a fully functional module can be swapped with the degraded one, which could in turn be repaired off-line. Each motor is also equipped with its own fail-safe brake so that the manipulator can be powered down in any configuration.

SENSORS

Each joint provides absolute joint position, derived joint velocity, and torque sensing for servo control, as well as motor winding temperature sensing for safety monitoring.

The joint position sensing scheme uses two sensors for each joint. The current manipulator design uses a potentiometer and a brushless resolver. Both are geared to the joint output using precision anti-backlash gears. These devices operate in a "two-speed" mode, providing much higher resolution than can be obtained from either one individually. In addition, the dual sensing scheme provides recovery from single point failures. If the resolver fails, the potentiometer can provide joint position feedback, with reduced servo bandwidth to compensate for the reduced resolution. If the potentiometer fails, the joint can continue to operate normally until the next power cycle, when the absolute joint position must be determined.

The output member of each joint includes special structures instrumented with strain gauges such that joint axis torque measurements can be obtained. The strain gauge signals are amplified using a full bridge amplifier circuit that resides within the joint module. The joint torque information can be used for advanced control techniques such as force reflection or joint torque servoing.

The manipulator wire harness includes dedicated lines to support end effectors or sensors. A single D-connector at the toolplate provides 40 wires. These lines originate in an electronics enclosure, where they can interface with end effector controllers or sensor processing electronics.

CONTROL SYSTEM

The manipulator is controlled by a hierarchical multiprocessor controller that uses advanced control algorithms for high level dexterous motion control and low level joint servo control. The control computer is VME bus-based. It uses three 680x0 family processors along with various data acquisition, memory, and communications devices. The embedded control, or "target" system is linked via Ethernet to a Sun workstation, which serves as the host computer for the graphical user interface. All of the system software is written in the C language and executes on the target system under the VxWorks real time operating system. As future generations of higher-performance hardware and new control techniques become available, this architecture simplifies the evolution process and lengthens the system's technological life.

Manipulator control algorithms include an endpoint control algorithm for task space commands and redundancy resolution, and joint level servo algorithms for tracking the high level commands. The endpoint control algorithm transforms workspace setpoints into joint angle commands while resolving the single redundant degree of freedom. The algorithm provides a "position to position" solution for the joint angle commands, rather than a "rate to rate" pseudoinverse solution with its inherent drawbacks [6]. Options for using the redundancy include manipulator configuration optimization, joint limit avoidance, and singularity avoidance. These criteria may be balanced against one another by setting simple numerical weights that are available through the user interface. Configuration optimization enables the user to specify the manipulator "pose" as well as its tool position and orientation. For example, the user could manipulate in a constrained area by commanding an "elbow down" pose when reaching under an obstacle, or by requiring that the "arm plane" (formed by the upper arm and forearm links) remain horizontal while reaching through a horizontal opening. By specifying the manipulator configuration in addition to the endpoint position, the manipulator motion has the "cyclicity" property - closed endpoint/configuration trajectories have corresponding closed joint space trajectories [6].

Joint level servo algorithms employ a combination of conventional linear control techniques along with advanced nonlinear methods. Feedback loop gains are parameterized by effective joint inertias, which helps to maintain constant joint servo bandwidth throughout the workspace. Additional feedforward terms compensate for gravity loading and manipulator inertia, reducing the required feedback loop bandwidth.

The control system includes safety features to protect the manipulator from error conditions and hardware failures. A "watchdog" process constantly checks sensor signals, algorithm calculations, and computer hardware and stops manipulator motion if it detects any error conditions. The inherent fault tolerance of the actuator and sensor systems makes it unlikely that common types of hardware failures will leave the manipulator marooned.

The graphical user interface enables the user to specify motion trajectories, set algorithm parameters, and determine the manipulator status. As the user selects various operating modes, such as endpoint motion, single joint motion, or playback, different control panels are displayed. These panels enable the user to set motion parameters as well as start and stop the manipulator. For example, the user can define a trajectory using the current manipulator position as the origin. Endpoint trajectories can be defined with respect to different coordinate frames, such as the base or toolplate frames. The user can command both new endpoint or joint trajectories, or replay trajectories stored in a file. The control system enables various algorithm parameters to be changed "on the fly" so that the redundancy resolution criteria and manipulator response characteristics can be modified in real time.

One of the software design's most important features is the interface definition which simplifies modifications and extensions to the control system. Critical data, such as sensor values and motion commands, is accessible from shared memory via simple function calls, greatly simplifying the process of adding new software modules which must use this data. While a commercial end user may not require such low level interaction with the control system, a researcher investigating advanced control algorithms would demand such access. The shared memory architecture provides this access. In fact, the researcher could write modules of C code to implement his algorithms, making the appropriate function calls to access sensor data such as joint angles and rates, compile and link these modules with the rest of the control system software, and evaluate the results using the actual manipulator. As an example, bilateral teleoperator control with a force reflecting master controller could be added by writing a software module that simply

makes the appropriate function calls to place master commands into the manipulator endpoint command memory space In the same way, endpoint forces could be calculated from the sensed joint torques and transmitted back to the master's computing node.

Odetics has developed and continues to develop advanced control techniques, algorithms, and software for manipulators. In addition to the endpoint and configuration optimization algorithms developed with this manipulator, the company has previously developed algorithms for dual coordinated manipulator control with collision avoidance capability. Path planning and trajectory generation algorithms are currently being developed for the Dexterous Manipulator The path planning algorithms will find the shortest path around obstacles in the manipulator workspace to a goal position for the manipulator end effector. The trajectory generations algorithms use a potential field approach to guide the end effector along this path while simultaneously avoiding collisions between the end effector, the links of the manipulator, and obstacles in the manipulator workspace. The resulting trajectory can be converted to joint angle commands and input to the joint servo control algorithms.

APPLICATIONS

This manipulator is targeted to address applications which, in addition to dexterity, require the strength and force control of a hydraulic manipulator, but for which hydraulic systems are impractical or impossible. A key member of this class is space manipulators. Many of the initial implementation's sizing and fault tolerance characteristics are chosen to be consistent with space applications. Some potential space applications include satellite servicing operations and truss assembly. A second member of this class is handling of hazardous materials in unstructured terrestrial environments. The manipulator workspace and payload have been chosen sufficiently robust to support both single and dual arm applications. Specific applications in this area include site characterization, radioactive waste handling, and hazardous materials disposal. The manipulator is suitable for tasks such as dexterously positioning and pointing a sensor or handling and transporting cannisters weighing up to 50 pounds.

As industry continues to shift towards a batch-oriented, flexible manufacturing paradigm, the ability to modify production equipment quickly and inexpensively will become increasingly important. As discussed earlier, the structure of this system's hardware and software are highly modular, and will support a wide range of applications that differ substantially from the present seven degree of freedom manipulator. The benefits of system fault tolerance, high precision, and high torque capability can be realized in applications with significantly fewer degrees of freedom, or conversely, in even more complex kinematic configurations. For example, an automated manufacturing operation could require that a tool be pointed accurately under high load, perhaps in an environment with coolant spray or chips. An alternative to custom-designed "hard automation" could be a two module pointing unit. This unit could readily provide the required motion control and load capability, with the added benefit of fault tolerance and very fast change-out if a failure occurs.

COMPARISON TO ROBOTICS RESEARCH MANIPULATORS

As mentioned in the introduction, the commercially available manipulator closest in configuration and performance to the Dexterous Manipulator comes from the Robotics Research Corp. line of modular dexterous manipulators. The K/B-1207 model has a 47.25 inch reach, 20 lb continuous duty payload, and 160 lb. weight, compared to the Dexterous Manipulator's 55 inch reach, 25 lb. continuous payload, and 150 lb. weight. Similarities between the manipulators include electric actuation, joint torque sensing/control, and hierarchical control system implementation. There are also several important differences between the manipulators.

Perhaps the most tangible difference between the systems is that the Robotics Research manipulators have been marketed as commercial products: units have been delivered and installed at various sites. The Dexterous Manipulator described in this paper is a prototype unit. There are important design and implementation differences as well:

1. The most salient design difference is the Dexterous Manipulator's actuator topology that provides fault tolerant sensing and actuation, enhancing the benefits gained from modularity. In particular, fault tolerance would appear mandatory in space and hazardous terrestrial applications where human intervention is extremely undesirable.
2. The actuator reduction methods are different - the Robotics Research joint modules use harmonic drives, while the Dexterous Manipulator uses spur gears exclusively. This reduction method combined with the unique actuator topology provides zero backlash over the life of the mechanism, with no adjustments required. The reduction method also provides somewhat higher forward and backdriving efficiency than harmonic drives.
3. The Dexterous Manipulator module interfaces completely enclose and protect the internal joint components, leaving only D connectors (which are mounted in the joint module flanges) exposed. There are no loose wires to connect during the mating process. These design features contribute to achieving "painless" modularity.
4. The 7 degree of freedom manipulator configurations are somewhat different kinematically. In addition, with its 6.5 inch diameter shoulder module and very smooth exterior lines, the Dexterous Manipulator is extremely compact.
5. The controller architecture and hardware are different - The Robotics Research controller uses both analog and digital control loops. The control computer is Multibus based and uses Intel 80X86 processors. The Dexterous Manipulator controller uses all digital control loops. Its computer is VME bus based and uses Motorola 680X0 processors. The completely digital implementation, VME hardware, and open architecture enhance the control system's adaptability and make it a particularly attractive research tool.

SUMMARY

Odetics has developed a modular Dexterous Manipulator that can be reconfigured for various tasks. The control system algorithms and software have been developed concurrently, yielding an integrated system. Although not currently space qualified, the system design and performance features make it suitable for both space and terrestrial applications. Potential space applications include satellite servicing and refueling, space truss assembly, and Space Exploration Initiative support operations. Terrestrial applications for the 7 degree of freedom manipulator configuration include hazardous material handling in unstructured environments, while reduced degree of freedom configurations could be useful in situations requiring high torque, fault tolerant, and accurate motion components for industrial processes.

In addition to the manipulator, the company is currently developing complementary systems, such as multi-fingered end effectors, a 7 degree of freedom exoskeleton master controller, and advanced control techniques for path planning and collision avoidance. These components can be integrated into complete systems that have the potential to greatly extend the capability envelope of robotic manipulators.

ACKNOWLEDGMENT

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Aeronautics and Space Administration.

REFERENCES

[1] Andary, J.F., Hewitt, D.R., Hinkal, S.W., "The Flight Telerobotic Servicer Tinman Concept: System Design Drivers and Task Analysis", Proceedings of the NASA Conference on Space Telerobotics", January, 1989, pp. 447-471

[2] Herndon, J.N., Babcock, S.M., Butler, P.L., Costello, H.M., Glassell, R.L., Kress, R.L., Kuban, D.P., Rowe, J.C., Williams, D.M, "The Laboratory Telerobotic Manipulator Program", Proceedings of the NASA Conference on Space Telerobotics", January, 1989, pp. 385-393

[3] Wu, E., Diftler, M., Hwang, J., "A Fault Tolerant Joint Drive System for the Space Shuttle Remote Manipulator System", Proceedings of the 1991 IEEE International Conference on Robotics and Automation, Sacramento, CA, April, 1991, pp. 2504-2509

[4] Jacobsen, S.C., Smith, S.M., Backman, D.K., Iversen, E.K., "High Performance, High Dexterity, Force Reflective Teleoperator II", Proceedings of the Fourth ANS Topical Meeting on Robotics and Remote Systems, Albuquerque, NM, February 24-27, 1991, pp. 393-402

[5] Karlen., J.P., Thompson, J.M., Farrell, J.D., "Design and Control of Modular, Kinematically Redundant Manipulators", Second AIAA / NASA / USAF Symposium on Automation, Robotics, and Advanced Computing for the National Space Program, March 9-11, 1987, Arlington, VA

[6] Baillieul, J., "Kinematic Programming Alternatives For Redundant Manipulators", IEEE International Conference on Robotics and Automation, St. Louis, MO, 1985, pp. 722-728

Figure 1 **The Odetics Dexterous Manipulator**

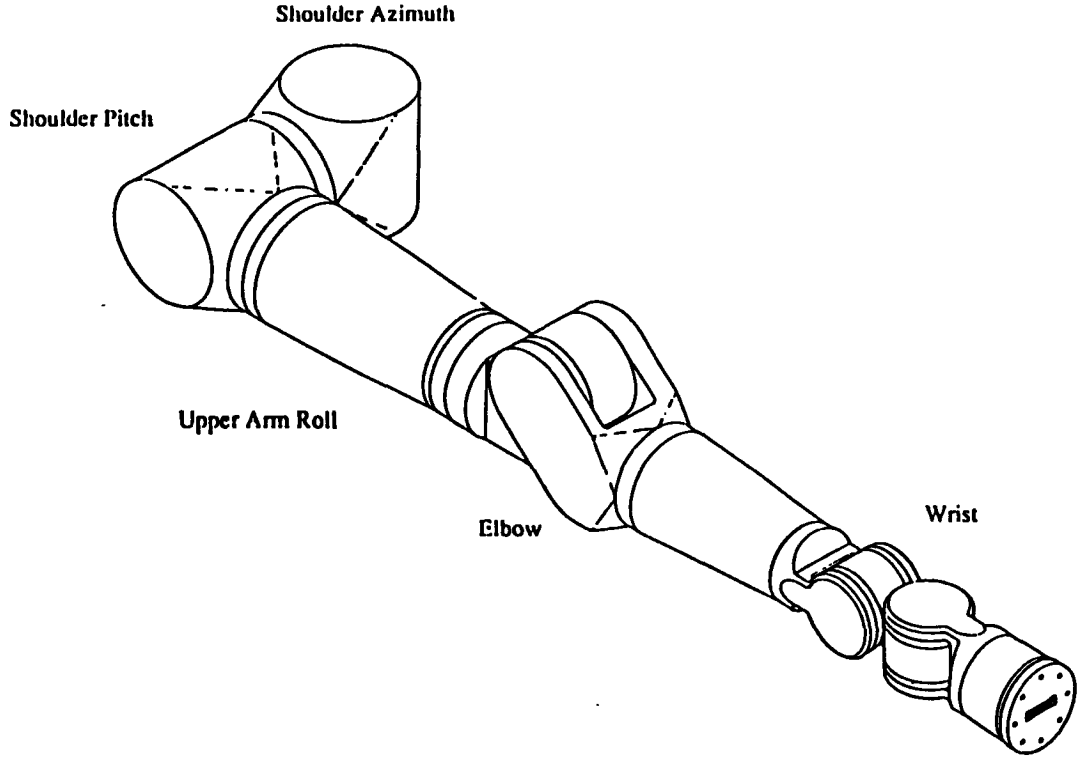
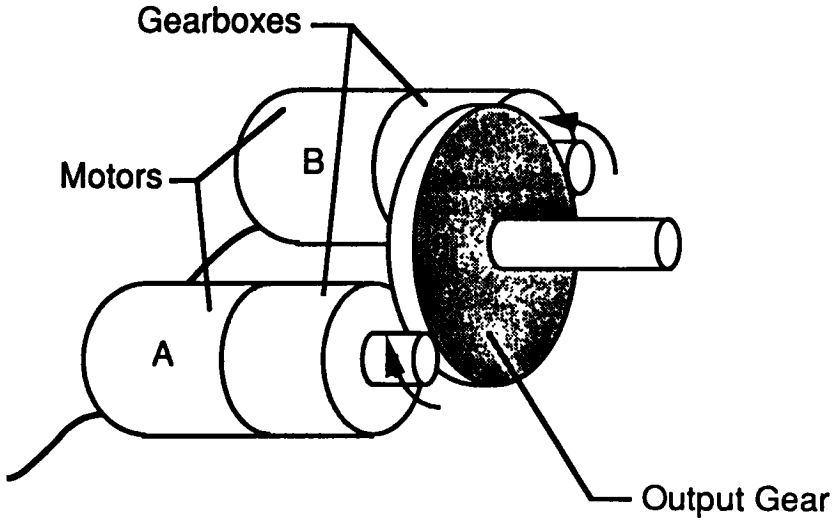


Figure 2 **Dual Motor Drive Concept**



N92-22683

REAL-TIME, INTERACTIVE, VISUALLY UPDATED SIMULATOR SYSTEM FOR TELEPRESENCE

Frederick S. Schebor and Jerry L. Turney
KMS Inc.
Ann Arbor, Michigan 48106-1567

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

Time delays and limited sensory feedback of remote telerobotic systems tend to disorient teleoperators and dramatically decrease the operator's performance. To remove the effects of time delays, we have designed and developed key components of a prototype forward simulation subsystem, the Global-Local Environment Telerobotic Simulator (GLETS) that buffers the operator from the remote task. GLETS totally immerses an operator in a real-time, interactive, simulated, visually updated artificial environment of the remote telerobotic site. Using GLETS, the operator will, in effect, enter into a telerobotic virtual reality and can easily form a gestalt of the virtual "local site" that matches the operator's normal interactions with the remote site. In addition to use in space based telerobotics, GLETS, due to its extendable architecture, can also be used in other teleoperational environments such as toxic material handling, construction and undersea exploration.

INTRODUCTION

To reduce the number of dangerous extravehicular activities required of astronauts, NASA will need telerobotic and autonomous robotic systems [1], [2], [3]. These intelligent robotic systems will build structures in orbit, perform repetitive manufacturing tasks and conduct repair, resupply, and servicing. Since the costs, risks and the required logistical support for astronauts to accomplish these tasks are very high, the need exists to perform as many functions as possible autonomously under ground based supervisory control. Because autonomous robots are currently beyond the state-of- the-art, NASA must initially use telerobotic or shared telerobotic/autonomous systems employing human operators to direct remote robotic manipulators. A major problem with these systems, however, stem from the delays inherent in transmitting data (~6 sec) to and from the remote telerobotic site. Because of these delays, conventional teleoperation using instantaneous feedback is not possible. Instead, the teleoperator and the teleoperation must operate in a decoupled manner. Unfortunately, teleoperation with time delays causes enormous operator fatigue and, as a result, can cause a dramatic increase in the number of potentially dangerous operator errors [4]. To compensate for this, NASA's Jet Propulsion Lab (JPL) has developed concepts of shared control [5] and "phantom" telerobotic systems [4]. The latter provide real-time visual feedback to an operator by predicting the motion of the remote telerobot and by graphically simulating the telerobotic motion and presenting this data in real-time as a response to the operator's inputs. The problem with this approach is that simulations can never perfectly account for the physics of the remote site, and, as a consequence, the simulator will eventually fail to accurately predict the state of the remote site. Therefore, a mechanism is required to continuously monitor the state of the remote site and feedback differences to allow the simulator to be operated in a closed loop.

To simplify teleoperation, KMS has designed a Global-Local Environment Telerobotic Simulator, GLETS. The GLETS design (Fig. 1) eliminates the lag between operator input and visual feedback response by interposing a simulator between the operator and the remote site. GLETS:

- simulates a remote manipulator and its local environment,
- continually updates the simulated environment using remote visual sensors,
- manipulates the simulated environment with an easy-to-use, gesturally and voice controlled operator interface, which provides rich visual and audio cues that can be easily interpreted by the operator,
- uses CAD databases for the simulated robot environment and robot descriptions,
- provides a flexible object-oriented software architecture with underlying standard robotics algorithmic support, which results in software that is reusable, extensible, reliable, and portable.

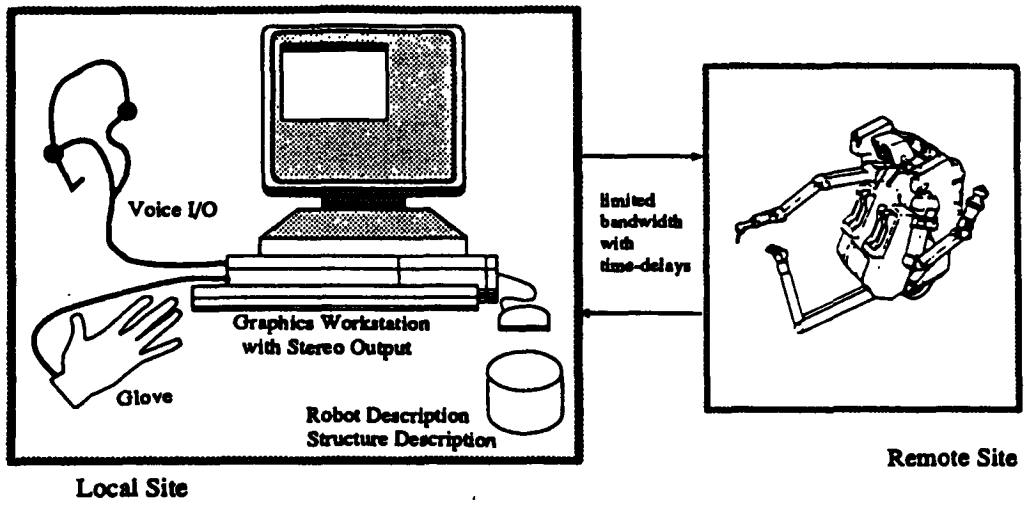


FIG. 1. By inserting GLETS between the operator and the remote manipulator, the simulator can shield the operator from problems of delays in the manipulator, interpreting the sensory input from the manipulators environment, and maintaining a detailed mental model of the manipulator's environment

OPERATOR INTERFACE

The effectiveness of a telerobotic system as a tool depends largely on the way the system is interfaced with the operator [6]. Due to the poor quality of telerobotic interfaces in the past, teleoperator training has been extremely expensive, and attempts have been made to identify favorable operator characteristics for shorter training time and safer system performance [7].

Current telerobotic simulators suffer from similar problems and many interfaces, such as those of ROBCAD and DENEB, concentrate on CAD tools for building up robot workcells. Because one must interact with the simulated robot through a 2-D graphics screen, it is difficult to properly control the robot manipulator arm in its simulated three-dimensional environment. In addition, the flow of information and commands between the simulator and the operator are restricted by the current interface technologies of CRT terminals, keyboards, joysticks, and mice [8]. Users are required to structure and channel communication to fit the machine. The key to solving this problem is to make the interface to the simulator more "human-like" rather than requiring the operator to be more "machine-like."

These problems can be eliminated, in a sense, by transporting the operator to the remote site so that the task can be accomplished and monitored handily. GLETS implements this concept by totally immersing an operator in a real-time, interactive, visually updated, simulated environment of the remote telerobotic site. Using GLETS the operator will, in effect, be able to interact with a telerobotic virtual reality and to form a gestalt of the virtual "local telerobotic site" that matches the operator's interactions with the actual remote site. These capabilities are provided in the following ways.

Display

Visually, GLETS portrays the telerobotic environment as a three-dimensional, shaded, color graphics world (Fig. 2). The simulator provides stereo images by using polarized glasses that alternately block the left and right lens synchronized with a graphics display that portrays the simulated environment from two slightly different views. Shading, perspective, motion, and stereo are the strongest depth cues in the human visual system. By incorporating these cues, GLETS takes maximum advantage of the operator's visual system, thereby making the task of controlling the arm more natural to the operator, and providing a more realistic simulated environment.

The graphics screen of GLETS can be divided into a number of windows. The windows display the views that would be seen by a set of "pseudo-cameras" (Fig. 3). Pseudo-cameras are attachable to any point within the simulated environment. For example, the operator can attach a pseudo- camera to the gripper of the manipulator to provide a continuous view of what would be seen by a gripper-held camera. Or, for a particularly difficult task such

FIG. 2. GLETS output is in the form of 3-D stereoscopic, shaded surface graphics and synthesized speech.

as gripping an object from behind, the operator could attach a pseudo-camera to the grip position on the object and use it to guide the gripper to the proper position.

GLETS allows the feeling of depth, commonly referred to as "stereo disparity," to be changed. Disparity is the difference in view angle registered between two eyes. Therefore, in general, human stereo is useful only for objects that are in close proximity. GLETS allows this disparity to be controlled in order to allow distant objects within the simulated environment to have the same perception of depth as closer objects.

GLETS provides multiple light sources (pseudo-lights) that can be positioned by the operator's hand and attached anywhere in space (Fig. 4). This permits elimination of shadows in critical areas.

GLETS allows any object in the simulated environment to be made transparent in order to let internal or difficult-to-see objects be viewed. For example, if the operator is required to insert a bolt into a particular hole and it was

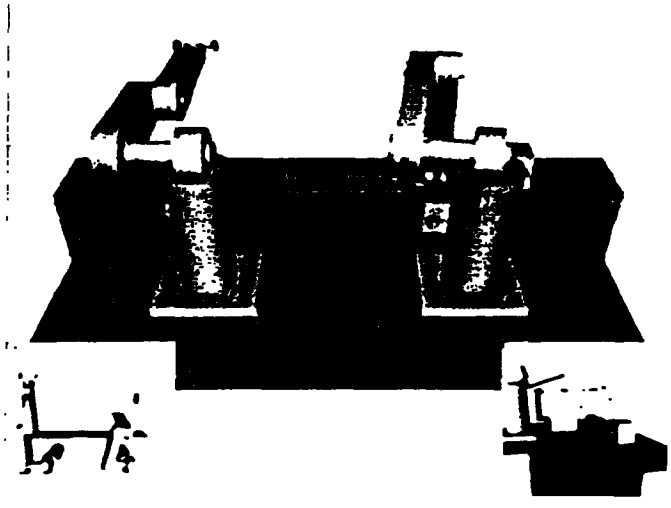


FIG. 3. The GLETS interface permits the operator to define and display multiple windows into the remote site

ORIGINAL PAGE IS
OF POOR QUALITY

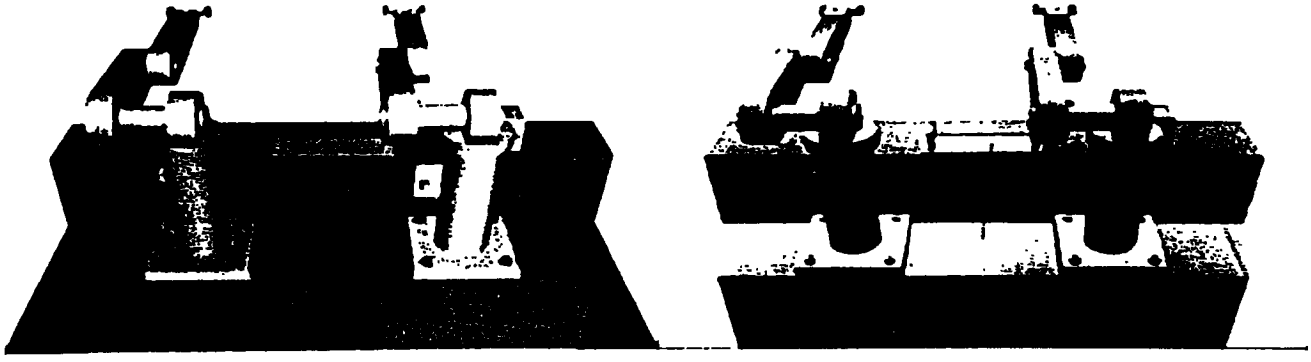


FIG. 4. The GLETS interface allows the operator to define multiple light sources (at infinity or local) at arbitrary locations.

not clear to the operator where the hole is located or how the hole is oriented, the entire object that contains the hole could be made transparent except for the inside surface of the hole in order to show the operator its position (Fig. 5).

Another visual attribute associated with objects in the simulated environment is the "glow" attribute which may be used to show the operator grip locations on an object or other special locations. These points are identified with a special color.

To test new control algorithms, GLETS provides on-screen information on the joint positions, velocities, accelerations, forces and torques (Fig. 6).

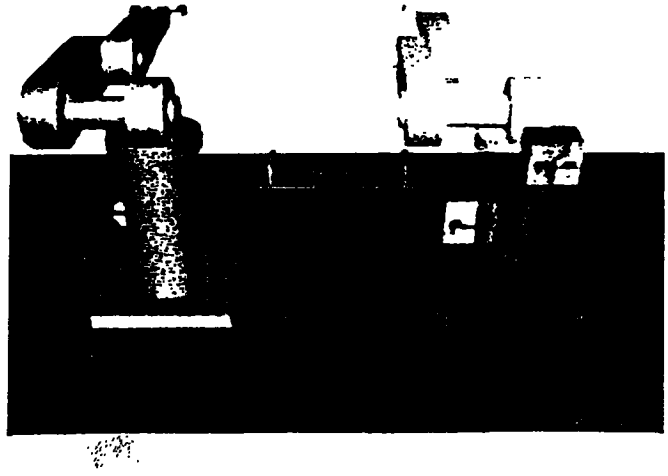


FIG. 5. With GLETS the operator can turn any objects at the remote site transparent in order to obtain a better relational understanding of the objects in the environment.

ORIGINAL PAGE IS
OF POOR QUALITY

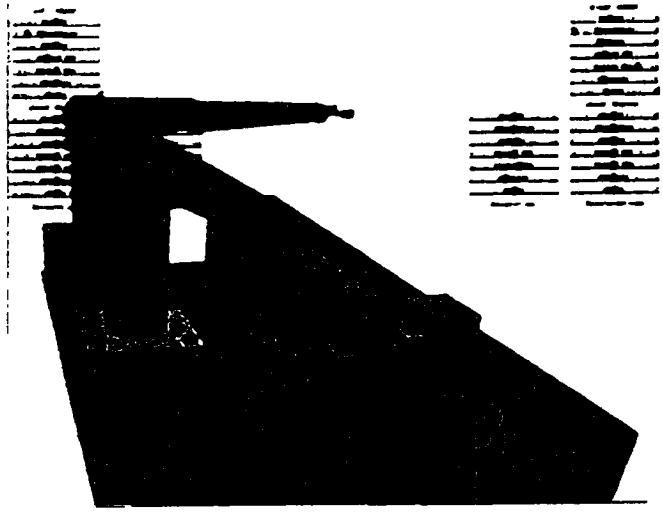


FIG. 6. The GLETS interface permits the operator to graphically view the physical parameters of robot motion.

Finally, in addition to shaded surfaces, GLETS can display any or all objects as wireframes (Fig. 7).

Computer-Linked Glove

A computer-linked glove [9] that allows the system to monitor the position, orientation and finger configuration of an operator's hand, permits the operator to use his hand to control the end effector. As the operator moves his hand, the system monitors the changes in position and orientation and passes these through reverse kinematics equations to determine the changes in the manipulator's joint angles necessary to match the hand position. The glove also permits the hand to be used as a "pick" device on virtual control panels, or through the use of gestures, to provide complex instructions to the telerobotic simulator, such as grasp, release, insert, extract, open, or close.

These complex instructions are implemented as a set of macros, or subprograms, that have been developed for

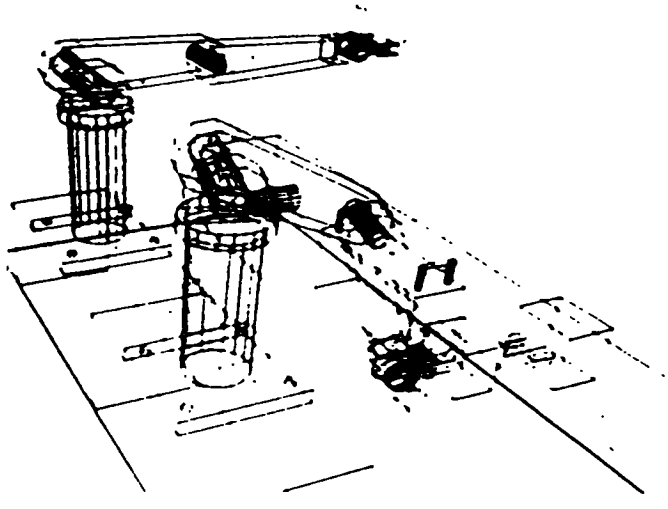


FIG. 7. The GLETS interface permits the operator to display objects as wireframes with hidden-line removal or with labeled object axis.

ORIGINAL PAGE IS
OF POOR QUALITY

special applications. Macros are used for two reasons. First, to simplify an operator's interaction with the environment. For example, an operator only need position a telemanipulator in the approximate vicinity of a handle, and execute the "grasp" macro either through a verbal command (described in the next sub-section) or through a gesture and the preprogrammed macro will "take over" and go through the necessary steps to grasp the handle. More importantly, however, macros provide a form of shared control [5] in which the telerobot can be controlled during free movement in the teleoperated mode, but must be controlled in autonomous mode during contact operations.

In addition to manipulating simulated objects, non-physical tasks, such as changing control parameters on a manipulator arm, are performed with object-like metaphors. To change a parameter, for example, the operator is provided with simulated slide bars and knobs that can be hand positioned using the glove.

Speech Synthesis and Voice Recognition

To provide input of more abstract commands than can be input through gesturing, and to notify the operator of simulation status, the GLETS system employs a speech synthesizer and a voice recognition unit. Voice input and output have produced productive gains in graphical workstation environments [10], [11], [12]. In GLETS, verbal commands are used to set up system states, execute macros and to receive status. This type of interface promotes easy, reliable control of the GLETS system. With little training an operator can quickly master the entire control of the system.

VISUAL UPDATE

Unless a robot simulator can maintain both a correct and current world model of its environment, it cannot function in a telerobotic application. To address this need, KMS has designed a software architecture and visual update subsystems that allow the simulator to interact with a remote vision system for update of the world model.

The update capability, in a sense, closes the teleoperational control loop by providing sensory feedback that can be used to minimize the discrepancies between the simulated and actual telerobotic environment. Without feedback, an open loop simulation can not completely characterize the remote environment. Therefore, discrepancies would start to appear that would be difficult for an operator to cope with. In particular, contact forces are notoriously difficult to model. If these are unknown, as is generally the case, then, an object could potentially slip in a gripper at the remote site and this slippage would not be incorporated in the telerobotic simulation. As a consequence the object might later collide with other objects in the environment, or the object might not be positioned correctly by the manipulator when it was used in a later insertion operation.

The visual update system serves two purposes:

1. to determine the minor discrepancies between the real and simulated world environments, and to update the simulator when these discrepancies become too large to tolerate,
2. to locate objects that are completely lost by the simulator.

Methods

To perform the first function, the visual update system makes real-time range (distance) data and image data measurements of the telerobotic environment. This could be done, for example, using a full-field image/range camera (such as described in the following sub-section). The range and image data can then be compared to the simulated range and image data generated by GLETS. JPL has used a similar approach [4] with single image video data.

In the same way that image data can be compared to simulated image data, range data can be compared to simulated range data, generated from the z-buffer data (used for hidden surface calculations) of the simulator, to determine discrepancies. Once a discrepancy is discovered, the visual update system determines the correct pose of an object using KMS developed vision algorithms, and uses this pose to relocate the object in the simulator. In GLETS we use an approach similar to JPL [4] to calibrate real and simulated image/range data. Specifically, points in the real environment (imaged by the range camera) and corresponding points in the simulated environment are hand selected. These are used to establish a mapping between the two environments.

When real and simulated objects are "out of sync" or when objects are completely lost by the simulator, GLETS employs an algorithm [13] that uses range (3-D) and/or image (2-D) data to locate 3-D objects. We refer to this algorithm as the RISER (Recognition by Iterative Spring Energy Reduction) algorithm. This robustness of this algorithm permits it to operate in cluttered environment, i.e., does not require that objects be completely visible.

Range-based RISER

In the range based RISER algorithm, 3-D model data representing objects that may appear in a scene, are matched to the range image of the scene. RISER requires that both the model and range data are represented as a collection of overlapping surface patches. Descriptors are then created by pairing up surface patches which are chosen from different regions of the surface. Descriptors of this type have a number of geometrical attributes, such as the angles between the normals at the center of each patch, the principle curvatures at each patch, etc., that allow them to be differentiated from each other.

To obtain a match, RISER sets up a set of pseudo forces between similar model and range descriptors. In effect, this is like connecting a 6-dimensional spring (3 translational and 3 rotational dimensions) between the descriptors, with a spring constant proportional to the descriptor's similarity (Fig. 8). The spring attempts to pull the corresponding model and range descriptors into alignment. To obtain convergence to a correct solution, the algorithm employs a robust statistical approach of reweighting the spring constants.

Image-based RISER

The first step of the image-based RISER algorithm is to construct shape representations of all objects in the image. As in the range-based algorithm, shape representations consist of a set of spatially local shape primitives described by various geometric attributes. The algorithm then uses an initial estimate of the object's pose in conjunction with a 3-D model of the object to predict the appearance of the edge contours. The predicted edge contours' shapes are represented by shape primitives in the same manner as the observed edge contours, permitting the observed shape primitives to be directly compared to the predicted shape primitives. The algorithm then connects pseudo-springs between the detected primitives and the predicted primitives, and as in the range-based RISER algorithm, the model is then freed and allowed to relax to the equilibrium pose under the influence of the spring forces.

The RISER pose determination approach has a number of advantages over conventional approaches. First, since it compares only pairs consisting of one image/range shape primitive and one predicted shape primitive, it avoids searching the exponentially large space of all possible image feature to model feature correspondences typical of many conventional approaches. Second, since RISER is comparing 2-D predicted contours to 2-D image/range contours, no assumptions about the nature of the surfaces of the objects are necessary.

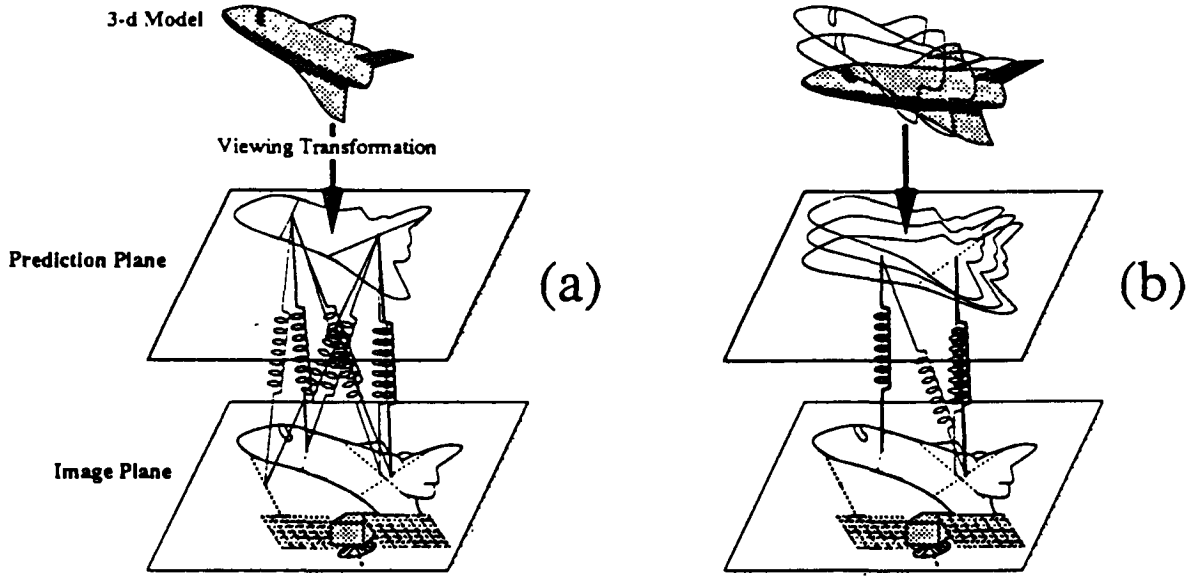


FIG. 8. RISER's advanced pose determination modules function, in effect, by connecting "springs" whose strengths are proportional to the similarity between features between primitives, a few of which are shown in (a). The model is then freed and allowed to relax into equilibrium, as in (b). The spring constants are then readjusted and the process continued until no further improvements occur.

Sensors

To address the need for visual tracking sensors, KMS has designed two different cameras that use structured lighting to extract range data. The RTRT camera, a Real-time Three-dimensional Range Tracking Sensor combines a rugged optical design with state-of-the-art phase-shift moiré interferometry [14]. The RTRT Camera generates moiré images, that encode the range of objects in the field of view, at an instant in time.

KMS is currently developing the other camera type for a "View-Generated Database" (VGD) project for NASA [15]. Our investigation into this problem has led KMS to a new type of structured lighting wherein phase shifted cosine patterns are projected using high-quality sinusoidal slides onto the surfaces of the scene using a CCD camera to image them. The phase values obtained by the CCD camera are used to triangulate the distance to a surface. We call this new technique SUrface Reconstruction by PHAse-shifted CosinEs (SURPHACE).

Using either of these image/range camera designs or other systems that are currently being developed, the GLETS visual update system is able to bring in image and range data that can be directly compared to calibrated, simulated data to determine where discrepancies have occurred.

REMOTE MANIPULATOR SIMULATION

The GLETS design also addresses the completeness of the simulation for the manipulator and its environment. Full simulation includes graphical, geometric, kinematic, dynamic, and control simulations [17]. Graphical simulation is concerned with simulating the appearance of the manipulator for human interpretation. Geometric simulation is concerned with determining the spatial relationships of the parts of the manipulator and, for example, determining if collisions have, or are about to, occur. Kinematic simulation is concerned with determining the joint positions, velocities, and accelerations necessary to yield desired manipulator trajectories. Dynamic simulation includes simulation of the torques and forces acting on the manipulator resulting from motions of the manipulator and any attached objects. Control simulation involves the determination of the proper torques, forces, and motor currents necessary to achieve a particular motion of the manipulator.

In the GLETS object-oriented environment (described in the next section) graphical, geometric, kinematic, dynamic, and control attributes are associated with the robot arm links and actuators. For example, if a new type of actuator motor were developed, the simulator could be quickly updated by developing a new subclass that would describe the object and would know how to model the dynamic and control behavior of the new motor.

To provide a graphics description of the arm, GLETS inputs an IGES formatted file of the manipulator from a Robot Descriptor File (RDF). Manipulator links then render themselves to provide a three-dimensional graphics mode. of the manipulator.

For the kinematic description of the arm, GLETS inputs data from the RDF, that contains the Denavit-Hartenburg kinematic parameters for the manipulator, the most common method of describing robot linkages. The ranges of the joint angles are also supplied, so that once the full kinematic and dynamic description of the robot is input, the robot is assembled and actuators will constrain themselves to move within the space of realizable joint positions. The dynamic description, also stored in the RDF, provides masses and inertias for all of the robot links as well as viscous friction coefficients for the joints. Using this data, GLETS is able to determine the forces and torques at any point on the manipulator.

KALI

The KALI subroutine library provides a flexible robot programming and control environment for coordinated multi-arm robots and has been used extensively by JPL [16]. Although KALI has been designed for coordinated motion of manipulators, the GLETS system will, a large fraction of the time, use the components of KALI that pertain to describing the kinematics and dynamics of individual manipulators.

For example, for a given trajectory (defined by the operator using the computer-linked glove in the operator interface) of the end effector of the manipulator, KALI routines are used to solve the manipulator arm equations to determine the joint angle trajectories. Although the operator's motion may be somewhat erratic, trajectories are approximated appropriately so that they consist of a sequence of straight line segments connected by smooth arcs. The acceleration is chosen to be zero when the manipulator is in the middle of a segment and a constant non-zero value during the transition between segments. Given the force and torque constraints of the manipulator, KALI subroutines are then used to limit the accelerations along the trajectory to values that will not violate these constraints.

SOFTWARE ARCHITECTURE

While the operator interface and the completeness of the simulation are critical to the success of the telerobotic simulator, also important is the simulator's flexibility and extendibility. An inflexible and unextendable simulator would quickly become obsolete as manipulator technology advances. Therefore maximum flexibility in the creation, updating, and customization of manipulator simulation models is important. Elements and parameters of existing models should be simple to change and update. The software architecture should include simple mechanisms for users to extend the simulator's software to handle wholly new manipulator designs as well as incorporating new kinematic, dynamic and control techniques.

One way of dealing with these problems in the GLETS design is to is use object-oriented programming [18], [19], [20] in describing the telerobotic environment. By using an object-oriented design GLETS has a software architecture based on the objects in the telerobotic environment rather than on the functions that are performed.

Functions tend to be more abstract and tied to a particular implementation. Objects, on the other hand, are more physical and provide a more meaningful metaphor for designing code. The GLETS system performs actions on certain objects within the defined environment, and can be used in a evolutionary implementation. The software used in simulating the telerobot can be viewed as an operational model of the world in which the robot exists. GLETS is organized around representations of the objects in the telerobotic environment so that the software structure reflects the physical structure of the telerobotic environment. Complexity of the GLETS system is controlled by creating, combining, and manipulating software objects instantiated from a set of generic software classes to perform the specific tasks of the system. For example, an orbital replacement unit, ORU, class could be defined as shown in Fig. 9.

CONCLUSION

The GLETS system provides an operator with sensors, operator interface, and software-architecture that uniformly treats a simulated telerobotic environment as being populated with a set of physical objects that can be viewed, manipulated, and coded in a highly intuitive fashion. This capability greatly simplifies training and use of the GLETS system.

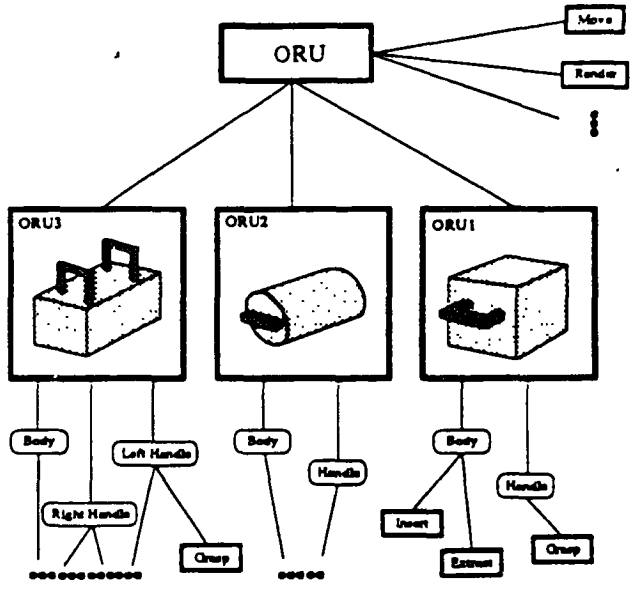


FIG. 9. GLETS object oriented software architecture supports the use of base and derived classes, a technique that greatly simplifies adding new objects to the system, because much of the properties of a new class of object can be inherited from its base class.

The most obvious use of GLETS will be in monitoring the state of a remote telerobot in a telerobotic operational loop in which severe time delays are expected. Although such conditions occur more severely between ground and space, they also occur in earth bound tasks such as undersea exploration, toxic waste management, remote inspection in nuclear and biological environments, and supervision in multiple automated production units. The GLETS system, because of its object oriented design, could be easily used in these areas as well as more generic applications such as evaluating new robotic algorithms as they evolve, facilitating robot configurations, mission planning/analysis, contingency planning, and error analysis.

ACKNOWLEDGMENTS

The authors wish to thank Arnold H.C. Chiu and Paul G. Gottschalk for their research efforts during this contract. The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This work was funded by NASA Small Business Innovative Research Contract NAS7-1074.

---

[1] M. L. Reiss, "Automation and robotics and the development of the space station - U.S. Congressional view," *Advances in Astronautical Sciences*, 60, 531, 1986.

[2] NASA Task Force, "Robotics for the United States space station," *Robotics*, 1, 205, 1985.

[3] R. M. Hord, *Handbook of Space Technology: Status and Projections*, CRC Press, 1985.

[4] A. K. Bejczy, W. S. Kim, and S. C. Venema, "The phantom robot: Predictive displays for teleoperation with time delays," *IEEE Conf. on Robotics and Automation*, 546, 1990.

[5] S. Hayati and S. T. Venkataraman, "Design and implementation of a robot control system with traded and shared control capability," *IEEE Conf. on Robotics and Automation*, 1310, 1989.

[6] C. S. Hartley, D. J. Cwynar, K. D. Garcia, R. A. Schein, "Capture of satellites having rotational motion," *Proc. of 30th Annual Meeting of the Human Factors Society*, 875, 1986.

[7] J. P. Yorchak, C. S. Hartley, and E. Hinman, "Characterization of good teleoperators: What aptitudes, interests, and experience correlate with measures of teleoperator performance," *Proc. of 29th Annual Meeting of the Human Factors Soc.*, 1135, 1985.

[8] M. B. Friedman, "Gestural control of robot end effectors," *Proc. SPIE Intelligent Robots Computer Vision*, 726, 500, 1986.

[9] T. G. Zimmerson, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "A hand gesture interface device," *Proc. of the SIGCHI Conf of the ACM*, 189, 1987.

[10] G. L. Martin, "The utility of speech input in user-computer interfaces," *Int'l J. Man-Machine Studies*, 30, 255, 1989.

[11] C. Schmandt, M. S. Ackerman, and D. Hindus, "Augmenting a window system with speech input," *IEEE Computer*, 23, 50, 1990.

[12] M. W. Salisbury, J. H. Hendrickson, T. L. Lammers, C. Fu and S. A. Moody, "Talk and draw: Bundling speech and graphics," *IEEE Computer*, 23, 59, 1990.

[13] P. G. Gottschalk, *Machine Recognition and Attitude Estimation of 3D Objects in Intensity Images*, Univ. Michigan Ph.D. Thesis, 1990.

[14] G. T.Reid, R. C. Rixon, and H. I. Messer, "Absolute and comparative measurements of three-dimensional shape by phase measuring moiré topography," *Opt. and Laser Tech.*, 16, 315,1984.

[15] J. L. Turney, "High performance view-generated database for world model definition and update," KMSF-U1923, NASA Contract NAS7-1009, 1987.

[16] P. Backes, S. Hayati, V. Hayward, and K. Tso, "The KALI multi-arm robot programming and control environment," *Proc. of NASA Conf. on Space Telerobotics* , 1989.

[17] T. N. Mudge and J. L. Turney, "Unifying robot arm control," *IEEE Trans. on Ind. Appl.*, 1A-20, 6, 1554, 1984.

[18] B. Meyer, *Object-oriented Software Construction*, Prentice Hall, 1988.

[19] R. A. Volz and T. N. Mudge, "Robots are (nothing more than) abstract data types," *Robotic Research: The Next Five Years and Beyond*, 1984.

[20] D. J. Miller and R. C. Lennox, "An object-oriented environment for robot system architectures," *IEEE Conf. on Robotics and Automation*, 352, 1990.

N92-22684

A HAZARD CONTROL SYSTEM FOR ROBOT MANIPULATORS

Ruth Chiang Carter
Goddard Space Flight Center
Greenbelt, MD 20771

Adrian Rad
Hernandez Engineering, Inc.
Greenbelt, MD 20770

ABSTRACT

Unlike the industrial robots in wide use around the world today, which generally perform various but limited repetitive tasks, a robot for space applications will be required to complete a variety of tasks in an uncertain, harsh environment. This fact presents unusual and highly difficult challenges to ensuring the safety of astronauts and keeping the equipment they depend on from becoming damaged. The application of system safety engineering to the design and development of the robot ensures that it will not become an instrument of harm or destruction to the space vehicle and its occupants. This paper describes the systematic approach being taken to control hazards that could result from introducing robotics technology in the space environment.

First, this paper will discuss system safety management and engineering principles, techniques, and requirements as they relate to Shuttle payload design and operation in general. The concepts of hazard, hazard category, and hazard control, as defined by the Shuttle payload safety requirements, will be explained.

Second, this paper will show how these general safety management and engineering principles are being implemented on an actual project. It will present an example of a hazard control system scheme for controlling one of the hazards identified for the Development Test Flight (DTF-1) of NASA's Flight Telerobotic Servicer, a teleoperated space robot. The paper will also discuss how these schemes can be applied to terrestrial robots as well. The same software monitoring and control approach will insure the safe operation of a slave manipulator under teleoperated or autonomous control in undersea, nuclear, or manufacturing applications where the manipulator is working in the vicinity of humans or critical hardware.

SYSTEM SAFETY ENGINEERING CONCEPTS

System safety is the overall management and engineering approach to the evaluation and reduction of risk in a system and its operation. In general, system safety activities include systematic identification of hazards, elimination of those hazards to the maximum extent possible, assessment of the residual risk inherent in the system or its operation, management review and acceptance of the risk, and documentation of the management decision and rationale in accepting the risk. In addition, some type of control for the hazards must be instituted.

In order to analyze the safety of any system properly, the system safety engineer must have a thorough knowledge of the system, subsystems, interfaces, functions, characteristics, intended use or operation, and the operational environment. It is therefore necessary for the system safety engineer to work closely with systems engineering, subsystem/component design engineering, mission operations, and support engineering personnel in developing a complete and accurate system description and operations scenario that can adequately support the identification of all potential hazards in the design and use of the system. Once an accurate system description is developed, the basic characteristics and functions of the system are defined, and preliminary operation scenarios are formulated, detailed hazard identification can begin.

The system safety discipline has established standard analysis methods and techniques for identifying hazards and categorizing their potential severity, including Preliminary Hazard Analyses, System/Subsystem Hazard Analyses, Operating Hazard Analyses, and Fault Tree Analyses. The analyses are prepared during system concept definition, design, and development in a cooperative and iterative process. The results can therefore

more effectively support management personnel in making decisions and in accepting risks; system safety personnel in assessing risks and identifying hazards; systems/subsystem, operations, and system safety engineering personnel in developing designs and operating procedures that eliminate and/or reduce hazards; support services personnel in conducting activities, such as configuration management and test and verification; and quality assurance personnel in documenting and verifying requirements and design criteria compliance.

SYSTEM SAFETY ENGINEERING IMPLEMENTED FOR DTF-1

The following paragraphs explain how system safety engineering was implemented for the (DTF-1). This example is intended to illustrate the system safety engineering process. In general, all of the information presented in this paper was gathered from DTF-1 design and safety engineering documents. Some information has been simplified and other information is not presented so as to facilitate a clearer presentation of the process.

System and Operation Description

The DTF-1 consists of the aft flight deck (AFD) element and the payload bay (PLB) element connected by two communication networks or buses. The AFD element, referred to as the workstation, is made up of a handcontroller, handcontroller electronics, and a control display console and assembly, including computer displays and video displays from cameras mounted on the manipulator and on other locations in the PLB, and a keyboard. The handcontroller is a device that is used to relay the motion, displacement, or relocation of the operator's hand as inputs to the control system. The handcontroller electronics transform the inputs into command signals to be sent to the manipulator in the PLB. The computer displays are the available manipulator/system parameters, sensor data, and manipulator position measurements, such as joint position, end-effector position, velocities, forces and torques, temperatures, and computer control system health and status. The AFD elements are provided electrical power through the power control and distribution unit (PCDU).

The PLB element, physically located in the Shuttle's cargo bay, includes the DTF-1 telerobot (TR), a TR control computer (TRCC), a TR redundant controller (TRRC), the power module and controller, and the payload controller (Figure 1). The TR consists of a telerobot body and a seven degree-of-freedom dexterous manipulator (Figure 2) with three computer controllers built into it--the shoulder controller, the upper arm controller, and the lower arm controller. These controllers, collectively referred to as joint controllers, provide joint position control loops and other control system functions. The TRCC provides the primary command and control function. The TRRC provides an independent safety checking and monitoring that includes collision avoidance and safety critical parameter monitoring. The power module and controller perform electrical power regulation and distribution control. The payload controller provides camera control, Shuttle-to-ground data downlink interface, and other support functions. Other PLB elements include an end-effector, a task panel, supporting structures, and other equipment. On the task panel are manipulation and articulation, partial disassembly, reassembly, removal, positioning, and reinstallation tasks.

The workstation bus connects the handcontroller electronics and control display console with the TRCC, the TRRC, the power module controller, and the payload controller. The TR bus connects the TRCC and the TRRC with the joint controllers.

As the operator at the AFD workstation positions and repositions the handcontroller, the handcontroller electronics transform the inputs into command signals that are sent to the TRCC. The TRCC then calculates the joint and end-effector motor commands needed to implement the commands; monitors manipulator position and velocity; and sends joint commands to the joint controllers. The joint controllers receive the commands, generate the needed joint motor currents, and monitor joint position and rotation against predefined limits. The joint position and torque sensors feed back data to the TRCC and the TRRC where the manipulator position and velocity are updated. This process is repeated continually throughout the teleoperation.

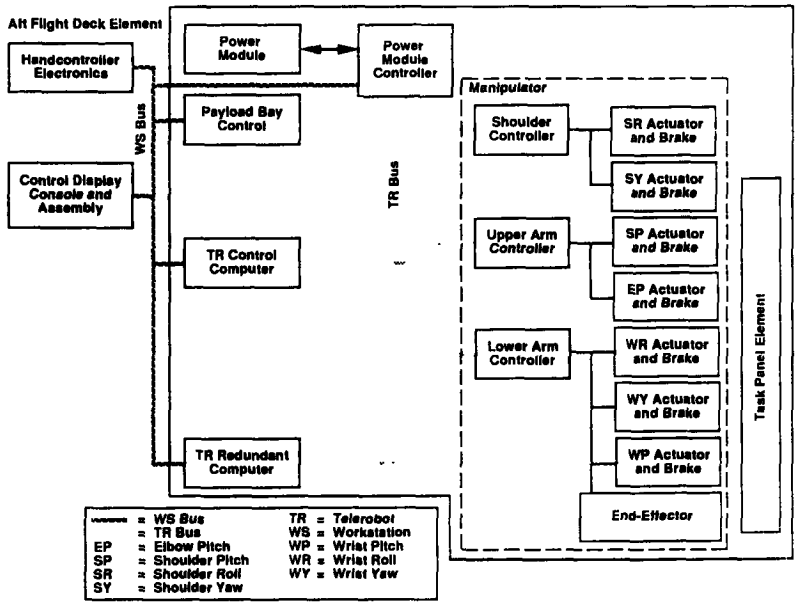


Figure 1. System Description Block Diagram

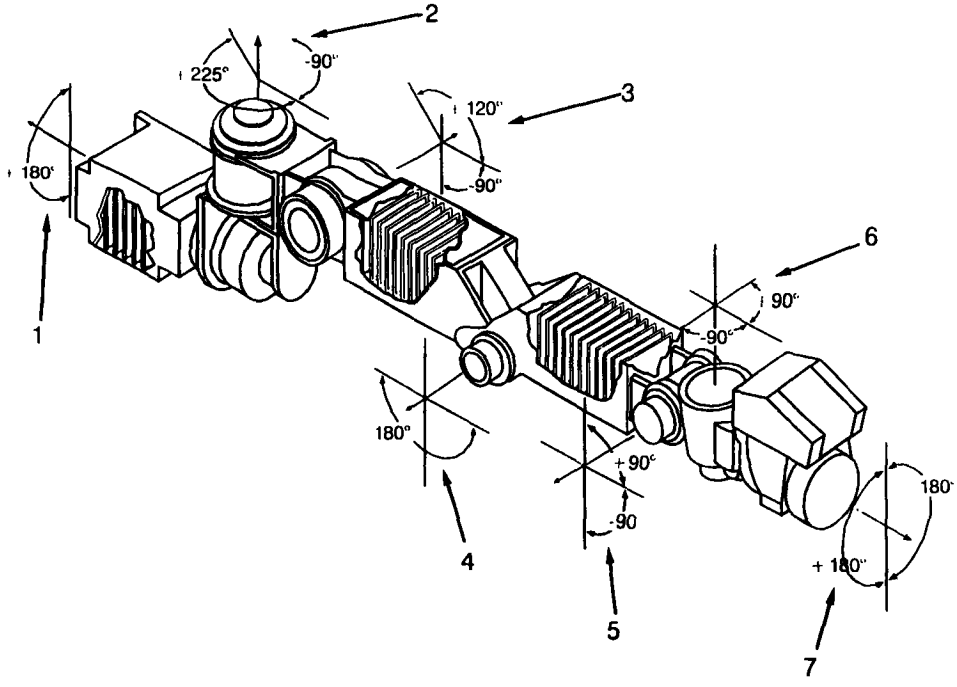


Figure 2. Manipulator

System Safety Requirements and Design Criteria Used

System safety engineering for DTF-1 was implemented with the system safety requirements and design criteria established by NASA's National Space Transportation System (NSTS) Program and defined in the Safety Policy and Requirements for Payloads Using the Space Transportation System, NSTS 1700.7B.
All hazards that could be eliminated were eliminated. The remaining hazards were controlled through safety devices, warning devices, or special procedures. The basic method for minimizing or controlling hazards, as prescribed by NSTS 1700.7B, is a failure tolerant design. The hazard severity category (i.e., critical or catastrophic) determines the failure tolerance design requirements. Critical hazards were controlled in a single-failure manner and catastrophic hazards were controlled in a two-failure tolerant manner.

Specific requirements and design criteria for subsystems, safety-related features, specific hazardous functions, and environmental compatibility were also included in NSTS 1700.7B to ensure the safety of the system design. They included structures, material compatibility, flammability, deployment and separation, contingency return and rapid safing, and others.

Hazard Identification

The potential hazards related to teleoperation of a dexterous, seven degree-of-freedom manipulator in orbit include collision or impact and excessive force and torque generated by the manipulator.

A catastrophic hazard could occur if the TR inadvertently collides into or otherwise impacts its surroundings during in-orbit operation. For example, if the motor and gear actuators generate excessive forces or torques at any of the seven TR joints or if excessive forces or torques are induced at the Cartesian level through the manipulator at the end-effector, then one or more of the manipulator joints may fracture or suffer other damage and the end-effector may generate excessive forces or torques on the task panel, surrounding structures, or equipment. Structural failure will cause the mission to be terminated. Also, it could severely compromise the ability to stow the manipulator safely and may create debris of sufficient size and mass to damage critical Shuttle equipment and prevent the Shuttle from returning home safely.

Another catastrophic hazard could occur if the manipulator is improperly positioned or is moving too fast and violates the predefined workspace. This situation along with other failures, such as loss of control of the manipulator's movement, could prevent closing of the payload bay door and return of the Shuttle.

Potential hazards for DTF-1, such as the ones just described, were controlled for with a two-failure tolerant hazard control system; i.e., three separate hazard control methods have been implemented. If two of the three methods fail, the third one will still control the hazard. The following paragraphs detail this system.

Hazard Control System

The hazard control system controls excessive forces or torques and prevents collision or impact through an integrated DTF-1 computer control, sensor, and feedback system. Forces and torques or position and velocity commands are limited, safety critical parameters are monitored, and redundant safety critical parameters are monitored. If safety limits are exceeded, electrical power is removed from the joint motors and the fail-safe brakes are engaged (Figure 3). This process is termed emergency shut down (ESD).

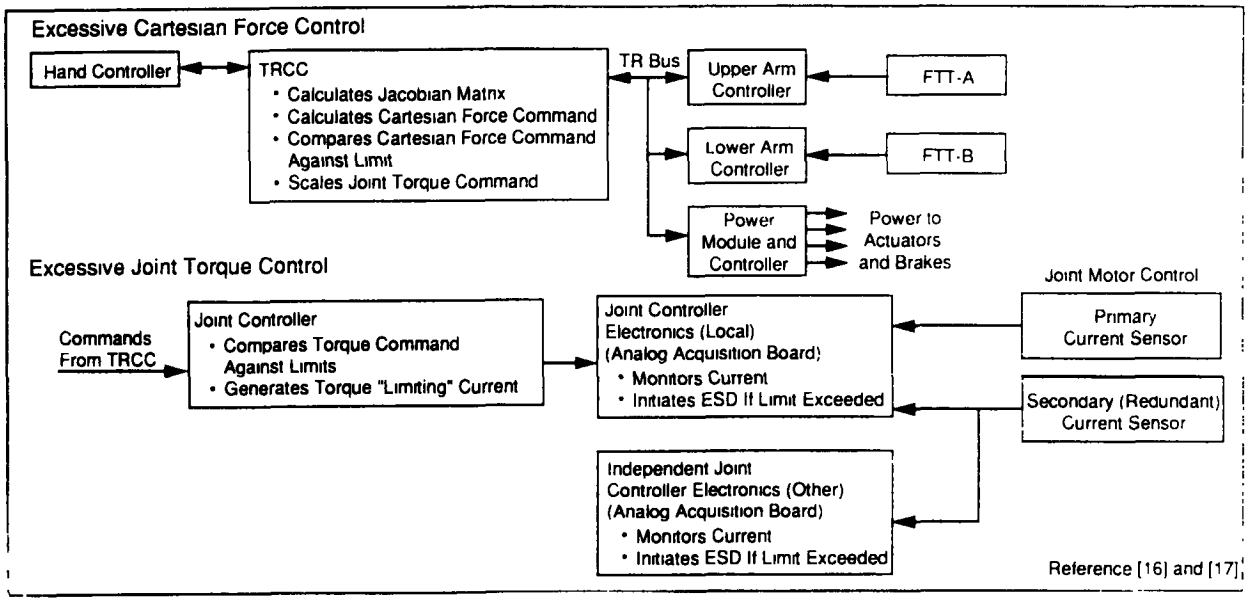


Figure 3. Excessive Force/Torque Hazard Control Scheme

Excessive Force and Torque Hazard Control

Force and torque commands are limited at the Cartesian level and at the joint level. Cartesian-level control limits the Cartesian position, velocity, and force at the manipulator and end-effector. Joint-level control prevents the motor and gear actuator assemblies from exceeding rated design torques. Joint-level control also provides a rapid response to joint "run away."

At the Cartesian level, hazard control is implemented as follows:

1) The TRCC calculates forces and torques for the end-effector from joint command inputs. Then the "commanded" forces and torques are compared against predefined safety limits. If the limits are exceeded, the inputs are automatically scaled down to acceptable levels before being passed onto the joint controllers.

2) Forces transmitted to the upper arm controller or to the lower arm controller by the force torque transducer (FTT) (located near the end-effector at the tool plate) are compared against predefined safety limits. If they are exceeded, an ESD is initiated.

At the joint level, hazard control is implemented as follows:

1) Current to the joint controllers is limited to that corresponding to the actuator-rated design torque.

2) The current in the primary and the secondary current sensors is monitored and compared to predefined safety limits. If they are exceeded, an ESD is initiated.

Collision and Impact Hazard Control

Collision and impact hazards are limited at the Cartesian and joint levels by controlling the manipulator's position and rate of motion (velocity). Limits on manipulator travel distance restrict the manipulator's position to within the predefined workspace. Limits on velocity control the force with which the manipulator impacts its surroundings. Cartesian-level controls limit the manipulator's movement in free space (work space) and joint-level controls limit individual joint movement in joint-space.

The distance at which the manipulator can be stopped is a function of the mass of the load the manipulator is carrying and the manipulator's velocity. The allowed velocity depends on the proximity of the manipulator to another object. Rate limits used in the servo algorithms are adjusted according to this distance. Violations of these limits require that the manipulator position commands be modified enough to slow its movement.

The boundary management and touch control (BMTC) system is used to control collision and impact hazards. This system sets imaginary (invisible) boundaries in and around the workspace, thus preventing unplanned contact which can result in collision and impact. Figure 4 shows how BMTC is implemented in the DTF-1's computer system. Figure 5 shows the functional scheme for BMTC. The four regions (X, A, B, and C) define the distance from geometric objects and are used to model each physical object. When the manipulator is in region C, normal free-space motion limits apply. In region B, the rate limits are reduced. At the outside of region A, a soft stop will occur unless or until the boundary is disabled. At the outside of region x, an ESD will occur unless the boundary is disabled. Region X and region A boundaries are disabled at the same time. A safety rate limit ESD can occur in regions A, B, or C if the operation rate limits are not applied.

Hazard Control System Assessment

Each safety feature of the DTF-1 hazard control system was analyzed at a system level to verify its effectiveness. This activity is important especially for a complicated system like the DTF-1 because failure tolerance is implemented only through the integration of all the safety features. A number of analyses were performed. However, only the fault tree analysis is used to illustrate the methodology and its usefulness in system evaluation.

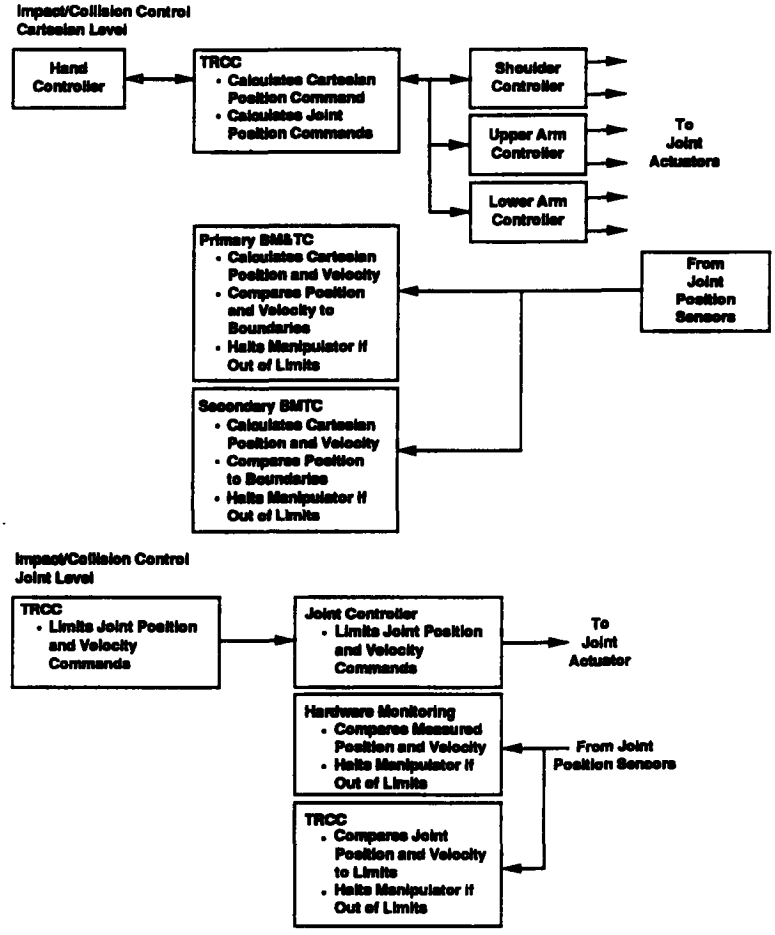
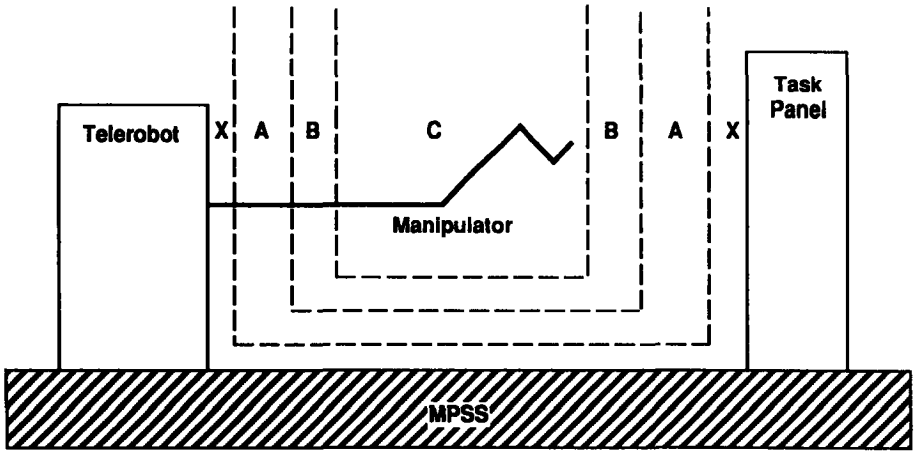


Figure 4. Collision/Impact Hazard Control Scheme



Region X: ESD Region
Region A: Automatic Soft Stop/Safe-Contact Rate Limited
Region B: Automatic Reduced Rate Limited
Region C: Normal Rate Limited

Figure 5. BMTC Functional Scheme

Fault tree analysis was adopted to evaluate hazard controls for excessive force/torque and collision/impact hazards. This analysis was chosen because of its top down approach. Through the analysis, it can be determined what components/functions must fail in order for a hazard to occur. If the results show that more than three components/functions must fail before a hazard can occur, then the implemented hazard control system is at least two failure tolerant. Figure 6 is a fault tree for the excessive force/torque hazard. It specifies the components/functions that directly contribute to the hazard.

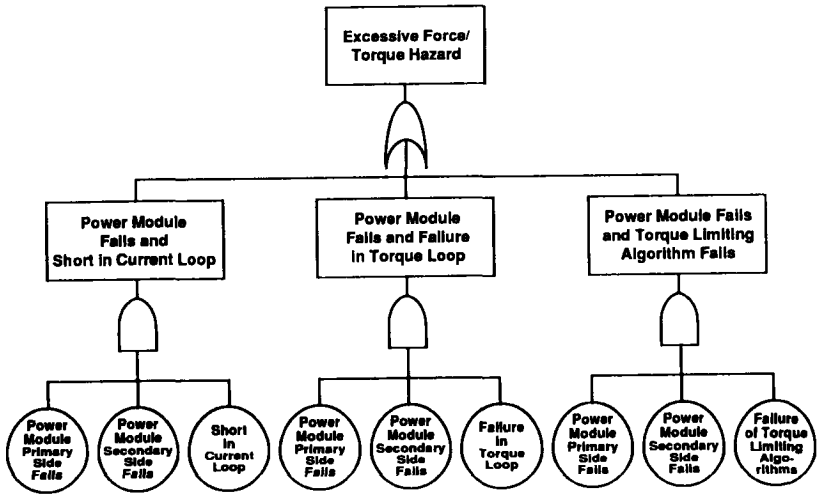


Figure 6. Fault Tree for Excessive Force/Torque

TERRESTRIAL APPLICATIONS

The DTF-1 hazard control system can be applied to any robotic and telerobotic system in which unplanned contact, impact, excessive forces and torques and manipulator motion control are of concern. It is applicable for terrestrial-based telerobotic and autonomous systems that must operate in varying environments, in confined workspaces, and near safety critical or hazardous equipment, such as those used in nuclear power plants, in under the sea operations, or on the factory floor.

In implementing the DTF-1 hazard control system, many of the design features common to all robotic systems, such as joint position sensors, force/torque sensors, and control computers and software have been used. As good system safety engineering dictates, the safety features would need to be customized and modified to the specific system design and operation for which they will be implemented. Equipment or components not available in the robotic or telerobotic system would not need to be incorporated.

CONCLUSION

The implementation of system safety engineering for the DTF-1 has been an iterative process. The initial concept was evaluated against safety requirements, and the system's fault tolerance was determined. In each of the subsystem areas, all possible hazardous events were hypothesized and their controls were evaluated. For example, two hypothesized hazard cases for which the analysis was completed were the excessive force/torque hazard and the collision/impact hazard. Once the possibility of these hazards actually occurring was confirmed, then designs and procedures were created to limit or prevent the hazardous commands from being generated.

The experience of implementing system safety engineering on the DTF-1 indicates that designing safety into the system early in the process maximizes the safety of the design and minimizes impacts to design complexity, development cost, and scheduling. Systems engineers, subsystem/components designers, and system safety

engineers are partners in creating an effective system safety scheme. Together, they will ensure that hazard control becomes integral to mission execution and a balance is achieved between a safe design and one that can accomplish maximum results.

ACKNOWLEDGEMENTS

Portions of this paper have been extracted from various project documents and technical meetings where FTS project staff provided valuable technical inputs. Specifically, Mr. Jim Andary, Systems Manager for the FTS Project, provided technical expertise and advice. In addition, the authors would like to acknowledge Ms. Barbara Hunt, without whose help completing this paper would not have been possible.

REFERENCES

[1] Flight Telerobotic Servicer Requirements Definition and Preliminary Design. SS-GSFC-0028, April, 1987.

[2] Flight Telerobotic Strawman Concept Engineering Report, SS-GSFC-0031, March 15, 1987.

[3] H. E. Roland, B. M. Moriarty, System Safety Engineering and Management, New York: John Wiley & Sons, Inc., 1983.

[4] J. F. Andary, S. W. Hinkal, and J. G. Watzin, "Design Concept for the Flight Telerobotic Servicer (FTS)" presented at the Second Annual Workshop on Space Operations Automation and Robotics (SOAR '88), Wright State University, Dayton, Ohio, July 20-23, 1988.

[5] J. F. Andary, D. R. Hewitt, and S. W. Hinkal, "The Flight Telerobotic Servicer Tinman Concept: System Design Drivers and Task Analysis," in the Proceedings of the NASA Conference on Space Telerobotics, vol. III, pp. 447-471, January 31, 1989.

[6] Safety Policy and Requirements for Payloads Using the Space Transportation System, NSTS 1700.7B, NASA Lyndon B. Johnson Space Center, 1989.

[7] System Safety, NASA Safety Manual, vol. 3, NHB 1700.1 (v3), April 11, 1984.

[8] System Safety for Orbital Flight Projects, NASA/Goddard Space Flight Center (GSFC) Management Instruction (GMI) 1700.3A, August 2, 1988.

[9] System Safety Program Requirements, Military Standard 882B (MIL-STD-882B), March 30, 1984.

[10] J. Hammack, A. Hernandez, S. Smith, "System Safety Training Course," Hernandez Engineering, Inc., Houston, Texas, December 1990.

[11] American National Standard for Industrial Robots and Robot Systems--Safety Requirements, American National Standards Institute/Robotic Industries Association (ANSI/RIA) R15.06, June 13, 1986.

[12] P. A. Lockner, P.D. Hancock, "Redundancy In Fault Tolerant Systems," Mechanical Engineering, pp. 76-83, May 1990.

[13] Hazard Analysis of the RMS (Robotic Arm) Simulation Using the Robot Safety Analysis, NASA/Goddard Space Flight Center (GSFC)/Code 205, December 22, 1988.

[14] Y. Sato and K. Inoue, "Safety Assessment of Human-Robot System: Hazard Identification Based on the Action-Change and Action-Chain Models," Bulletin of The Japanese Society of Mechanical Engineering (JSME), vol. 29, pp. 1351-1361, April 1986.

REFERENCES (Continued)

[15] Y. Sato, E. J. Henley, K. Inoue, "An Action-Chain Model for the Design of Hazard-Control System for Robots," IEEE Transactions on Reliability, vol. 39, pp. 151-157, June 1990.

[16] Space Station Flight Telerobotic Servicer (FTS) DTF-1 System Critical Design Review, 01-MD-02-DCR-01, prepared by Martin Marietta Corporation under contract to NASA/Goddard Space Flight Center (GSFC), September 25, 1990.

[17] Flight Telerobotic Servicer Phase C/D DTF-1 Phase 1 Payload Design and Flight Operations Safety Compliance Data Package, PA-32-05, prepared by Martin Marietta Corporation under contract to NASA/Goddard Space Flight Center (GSFC), May 4, 1990.

TEST AND MEASUREMENT

(Session C6/Room B1)

Wednesday December 4, 1991

- **Knowledge-Based Autonomous Test Engineer (KATE)**
- **Advanced Computed Tomography Inspection System (ACTIS)**
- **High-Resolution Ultrasonic Spectroscopy System for Nondestructive Evaluation**
- **Force Limited Vibration Testing**

N92-22685

KNOWLEDGE-BASED AUTONOMOUS TEST ENGINEER (KATE)

Carrie L. Parrish, Ph. D. and Barbara L. Brown
NASA Kennedy Space Center, DL-DSD-23
Kennedy Space Center, FL 32899

ABSTRACT

Mathematical models of system components have long been utilized to allow simulators to predict system behavior to various stimuli. Recent efforts to monitor, diagnose and control real-time systems using component models have experienced similar success. NASA at the Kennedy Space Center is continuing the development of a tool for implementing real-time knowledge-based diagnostic and control systems called KATE (Knowledge-based Autonomous Test Engineer). KATE is a model-based reasoning shell designed to provide autonomous control, monitoring, fault detection and diagnostics for complex engineering systems by applying its reasoning techniques to an exchangeable quantitative model describing the structure and function of the various system components and their systemic behavior.

INTRODUCTION

Conventional approaches to developing and maintaining diagnostic and control process software result in a time-consuming and costly effort. Furthermore, the resulting software may be incomplete and unable to handle situations that were unforeseen when the software was written. Significant advantages are obtained by systematically representing and using knowledge of a physical system's structure and function to reason about its health and proper functioning, a technique referred to as "model-based reasoning". The Artificial Intelligence Section, Engineering Development Directorate, at the Kennedy Space Center, employs the concept of model-based reasoning in the development of real-time knowledge-based diagnostic and control systems for ground launch operations. The project resulting from these efforts is called KATE (Knowledge-based Autonomous Test Engineer). KATE is being designed as a generic, model-based expert system shell, incorporating the engineer's reasoning about control and diagnosis of complex engineering systems in the form of general software algorithms. KATE further embodies concepts of sensor-based and model-driven monitoring and fault-location and performs control and redundancy management of process control systems through application of the generic algorithms to an exchangeable knowledge base, which describes the structure and function (i.e., mathematical model) of a specific domain. Refer to Figure 1 for a pictorial representation of the KATE System.

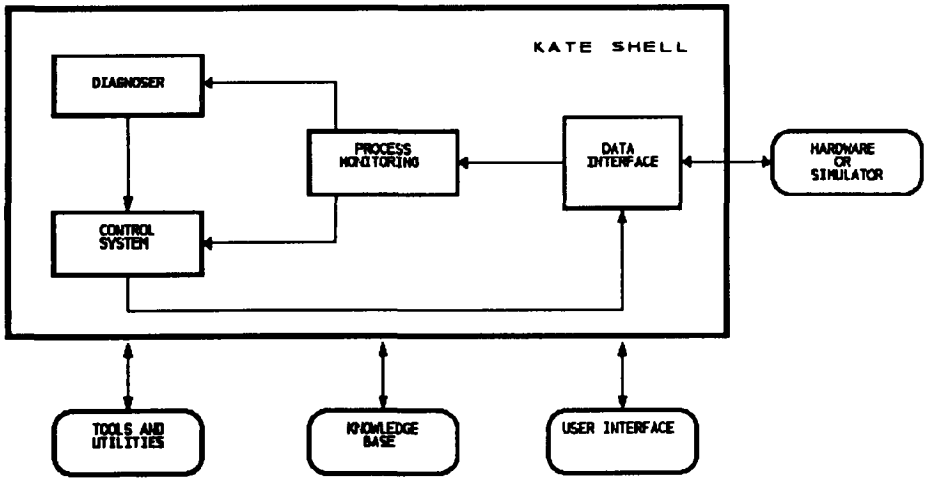


Figure 1. KATE Block Diagram

PRECEDING PAGE BLANK NOT FILMED

BACKGROUND

NASA at the Kennedy Space Center, first developed LES (LOX Expert System) in 1983 to perform fault detection and isolation for the process control system which handles the loading of Liquid Oxygen into the external tank of the Space Shuttle. In 1985, LES grew into the more generic and robust model-based system, KATE.

During fiscal year 1989, KATE was successfully demonstrated against a scale model of the Orbiter Modification and Refurbishment Facility's (OMRF) Environmental Control System (ECS). The OMRF ECS supplies conditioned air to four different compartments of the Orbiter while the Shuttle is being processed in the OMRF. The hardware model of the OMRF ECS contains a purge unit that supplies chilled air to four ducts which, in turn, supply air to the Orbiter. Each duct consists of a heater for maintaining a constant temperature and a motorized flow-control valve for controlling the flow rate. A failure panel has been added to allow for manual failing of various components during testing. An operator can request certain flow and temperature setpoints and KATE will respond by adjusting the valves and heater output to achieve the setpoints. KATE will also identify and respond to external inputs to the ECS, such as load changes. In addition to control, KATE monitors the system, identifying discrepant measurement values and assigning probable causes. In most cases, control of the system can continue, especially if the discrepancy is determined to be due to a measurement failure.

In February of 1990 implementation of the KATE Generic Control System (GCS) prototype was successfully integrated into the Generic Checkout System. The Generic Checkout System is comprised of a network of UNIX-based equipment connected to a physical model of a Shuttle water tanking system, referred to as the Red Wagon. KATE-GCS performed control and diagnosis of fastfill, chilldown, and other processes related to filling the Red Wagon external tank.

KATE KNOWLEDGE BASE

The knowledge base contains all information specific to the domain, i.e. the electro-pneumatic system being modeled. Generic component definitions as well as application specific data - classes of components, component names, tolerances, output-functions, inputs, delays, ranges, any information relevant to system operation, is stored here. The knowledge base can be thought of as a blue print of the system. When connected to hardware, the KATE shell utilizes this blue print to establish a nominal, working system model. This internal model can then be referred to as KATE monitors system health to detect anomalies, diagnose and recover from failures.

KATE's knowledge base has been developed through a combination of component modeling and mathematical modeling. That is, the knowledge base not only describes the system architecture, but is also the repository for functional information. System structure, or connectivity of system components, is represented primarily using data structures called inputs and outputs (described below). This information is used by KATE to discern control paths from command to measurement and is vital to gathering valid suspects during diagnosis of and recovery from failures. Mathematical equations of component output functions are also computed and stored in the knowledge base. Output functions describe how a component should behave based on its inputs. By representing structural and functional information, the dynamics of system's structure as well as its function can be made use of to perform monitoring, control and diagnosis. The knowledge base is separate from the KATE shell and therefore exchangeable. Conceivably, the shell might be used to monitor and control several hardware systems by the simple exchange of knowledge bases.

Knowledge Representation

The knowledge base is based on a hierarchy of components and is separated into 3 major levels: the top, mid, and instance level. System components are modeled from as high level a concept as a "thing" in the top level, down to specific individual line replaceable units, such as a bypass valve being operated in the field, in the instance level. The top level of the knowledge base provides the framework, or building blocks, for system modeling. Here, broad concepts and classes of components are defined. The mid level houses generic definitions of system components - butterfly valves, transducers, relays, flow meters, schematic pages, level sensors, circuit breakers, and fuses to name a few. The instance level contains domain specific information, such as calibrations for system components that override mid level default parameters. For example, the mid level generic definition of a valve may contain default operating ranges for the valve. This default may be overridden in the instance level

definition of an actual occurrence of a butterfly valve in the field of operation. Inheritance of frame information is accomplished via AIO and AKO slots (described below) during interpretation at run-time. The use of component inheritance provides a means by which any component can be trace backwards to its top level definition (or visa versa) and reduces the amount of information that must be stored in the knowledge base.

This object-oriented approach to knowledge representation greatly increases the ease of adding or removing system components. Should it become necessary to do so, a component may be deleted or "disconnected" from the knowledge base, connection points re-established, and the model is once again complete without the need for costly recoding that is necessary with conventional software design.

Pseudo Objects and External Influences

In an electro-mechanical system there exist components that are not actual "physical" components, but rather, engineering concepts that manifest themselves during system operation. The effects of these "non-physical" components are observable, and can be controlled and diagnosed if necessary. Early KATE developers found it useful to model these **pseudo-objects**, as they are called in the KATE knowledge base, for the proper functioning of the KATE shell software. In general, pseudo-objects represent information that has no corresponding sensor for measurement, but can be computed from other data. Examples of pseudo objects are pressures, temperatures, flow rates, and tank levels.

Manual valves and orifices are examples of **external influences**. When all system components are functioning according to expectations, manual operation of an external influence component might explain anomalous system behavior. Pseudo-objects and external influence object definitions, like any other objects, utilize inheritance and are defined at all levels of the knowledge base in the same fashion as actual existing, line replaceable components.

Knowledge Base Structure

A frame representation language has been employed for data representation. Frames contain "slots" of information, and utilities have been constructed to allow for manipulation of these slots during knowledge base development. Slots in a frame may be thought of as records in a data structure. Each frame contains a description of the object and its connection to other objects. At the instance level a frame definition can be found for each hardware component or pseudo object in the system. This statement however, is not meant to mislead the reader; for certain types of components, other frame definitions may be necessary to complete the system model.

KATE currently makes use of forty odd slots for modeling purposes. These slots contain information that applies to icon display, component location for camera tracking, or other information necessary for the KATE shell to effectively perform monitoring, control, and diagnosis. Selected knowledge base slots are described here for the reader's use in a later example of knowledge base construction.

NOMENCLATURE - Text string denoting the frame's object name and description.

AIO (An Instance Of) - Indicates the generic object type from which the component will inherit characteristics. For example, if A75064 is a pressure transducer, the AIO slot in A75064's instance level frame definition would read (aio PRESSURE-TRANSDUCER).

AKO (An Kind Of) - Indicates the class type of the object. For example, the AKO slot in the mid level frame definition of PRESSURE-TRANSDUCER would read (ako TRANSDUCER).

KINDS - Inverse slot link to AKO. List of all frames in the knowledge base that represent a class of object. For example, the KINDS slot in the top level definition of TRANSDUCER would read (kinds PRESSURE-TRANSDUCER...).

INSTANCES - Inverse slot link to AIO. List of all component names of a certain generic type existing in the field. For example, the INSTANCES slot in the generic component frame definition of PRESSURE-TRANSDUCER would read (instances A75064 ...).

INPUTS - list of all frame names of the object's inputs.

OUTPUTS - List of all frame names of the object's outputs.

CHANNEL - Indicates the Hardware Interface Module, card, and position number from which the measurement/command is polled/commanded.

COMPONENT-OF - Indicates a component's Hardware Interface Module and card number.

COMPONENT-POSITION - Indicates the position of the measurement or command on the polling card.

PHYSICAL-LOCATION - Text string denoting the components location in the field.

To better illustrate the method used to construct KATE's knowledge base from system schematics, a sample drawing, taken from an advanced electrical schematic is used (Figure 2). This drawing depicts a 3-way solenoid

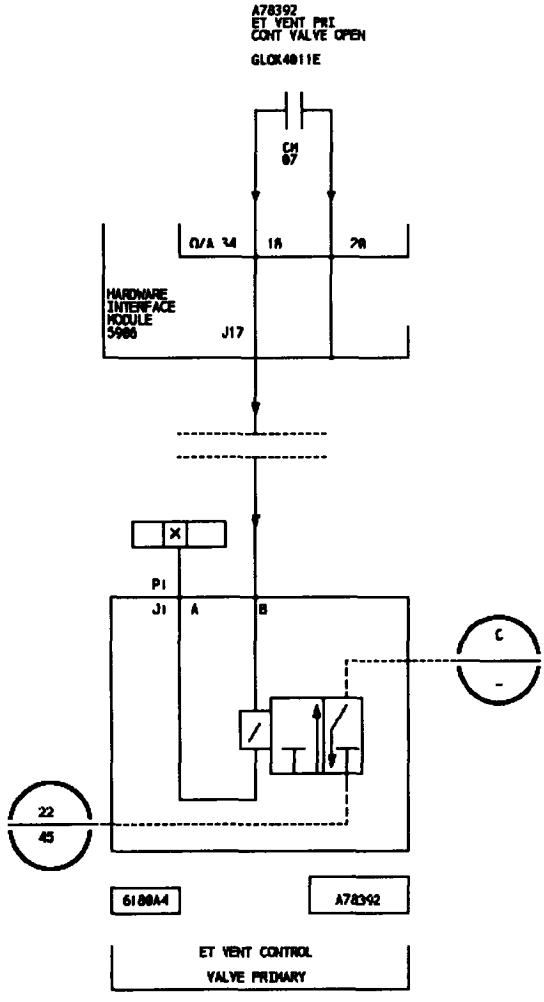


Figure 2.

TOP-LEVEL DEFINITION

```
(Deframe THING
(nomenclature "something")
(kinds SYSTEM
MEASUREMENT
COMMAND
DISCRETE OBJECT))

(Deframe DISCRETE-OBJECT
(NOMENCLATURE"a discrete object")
(also THING)
(kinds DISCRETE MEASUREMENT
DISCRETE COMMAND))

(Deframe COMMAND
NOMENCLATURE "a command")
(also THING)
(kinds DISCRETE-COMMAND
ANALOG-COMMAND))
```

MID-LEVEL DEFINITION

```
(Deframe DISCRETE-COMMAND
(nomenclature "a discrete command")
(also COMMAND
   DISCRETE OBJECT)
(instances GLOK4011E)
(output out))
```

INSTANCE-LEVEL DEFINITION

```
(Deframe GLOK4011E
(nomenclature "External Tank
 Control Valve Open Command")
(aio DISCRETE-COMMAND)
(outputs out (CH-5986-34-07 IN)
(channel CH-5986-34-07)
(component of CARD-5986-34)
(component position 07)
(physical-location External-Tank))

(Deframe A78392
(nomenclature "ET Vent Control Valve Primary")

(aio 3WAY-SOLENOID-VALVE-CLOSED
(inputs (coil (GLOK4011E- CONTACT out))
      (pressure-in1 (ATMOSPHERIC-PRESSURE out))
      (pressure-in2 (HE-SUPPLY-PRESSURE out)))
(outputs (pressure-out (A78394 pressure-in2)))
(physical-location ET)
```

Figure 3.

valve, **A78392**, and its associated primary open command, **GLOK4011E**. Instance level frame definitions for **A78392** and **GLOK4011E** are described (Figure 3), as well as the associated top and mid level frame definitions for a command. This example highlights the fact that there is not always a one-to-one correspondence of line replaceable component-to-instance level frame definition. While Figure 2 shows only a valve and its open command, in order to completely model **A78392** and **GLOK4011E**, the instance level file would also have to include frame definitions for **GLOK4011E**'s associated command measurement, command contact, and Hardware Interface Module card. (Note that top level object abstractions and the mid level generic component definition for a 3-way solenoid valve are not shown.)

KATE SHELL

Process Monitoring

Once the knowledge base (model) accurately reflects a physical system's behavior, monitoring is relatively straightforward. Process monitoring can be broken down into two main software modules, the Measurement System and the Constraint System. During normal operation, the Measurement System continuously polls command and sensor measurements through a hardware interface. This system then places all measurements that have changed by a user-defined "significant amount" on a queue for examination by the Constraint System.

The Constraint System processes any command or sensor measurement changes. If a command has changed, the new value is propagated through the model to generate expectation values for the system sensors. The sensor measurements placed on the queue by the Measurement System are then compared to the expectation values derived from the model. If all measurements are in agreement with the expected values, the assumption is made that the system is behaving properly, and monitoring continues. If, however, any sensor measurement differs from the expectation calculated using the model, it is assumed that the discrepancy is due to a physical system failure, and the Diagnoser is invoked to localize the fault.

Diagnosis

Introduction

The Diagnoser's task is to search for all possible failures within the system that can explain the current sensor readings. Reasoning from the sensor data, KATE uses a "violated expectations" approach to diagnosis which is a technique compatible with the intuitions of human diagnosticians. KATE currently assumes a single point of failure, however, the ability to diagnose to this point of failure is dependent upon the amount of sensor information available to KATE (i.e., the visibility into the system).

Diagnostic Algorithm

When the diagnoser is invoked due to a discrepancy, sufficient time is allotted for system stabilization. This time is given for the effects of a possible system failure to reach all of the system measurements (sensors) so that they may be used to aid in the diagnostic process. After system stabilization has occurred, any other discrepant measurements are gathered.

Initially, every component in the system which can be considered a potential "suspect" is placed on a suspect list. It is the diagnoser's task to reduce this list to a minimum number of suspects that can explain all current sensor information. These suspects are gathered using structural information found within the frames. Starting from the original discrepant measurement, all components that are structurally upstream from this sensor are considered suspect. The discrepant sensor is also placed on the list as a possibly faulty object. All other components are ignored (excluding structural faults like "bridges" or "short circuits" (2)) since objects not structurally upstream of this sensor cannot account for the discrepancy in its reading.

Once the suspects have been determined, all sensors which can be affected by these suspects are gathered. These sensors are also known as sibling sensors, and are used to aid in the effort of determining a suspect's innocence.

Further pruning of the suspect list is then attempted through the use of the functional information stored within the frames. Suspects which are of type "command" or "pseudo-object" are automatically cleared because it is assumed that these objects cannot fail. Furthermore, in the case that there exists more than one discrepant measurement, innocence can be established for the discrepant sensors based on the fact that a failed sensor cannot

cause another sensor to fail (refer to detailed description below).

A more sophisticated pruning method is then required to further localize the source of faulty system behavior. This method is based upon analysis-by-synthesis, a search for some fault that can explain what the sensors are showing, and has been labeled "The Full Consistency Algorithm" (3).

In general, the Diagnoser uses a "generate and test" paradigm. The algorithm involves calculating a hypothetical value for each suspect that would explain the discrepant measurement's value. In other words, the Diagnoser determines what state the suspect would have to be in to produce the discrepant measurement (generate). If the hypothetical state of the suspect consistent with the discrepant measurement cannot also account for the other system sensor readings (i.e., sibling sensors), then that suspect cannot be the cause of the failure and may be cleared from the suspect list (test). This hypothetical value is derived by performing symbolic inversion of the stored functional relationship between the discrepant measurement and the suspect. For further detail on the inversion process and a description of the inversion algorithm, refer to (4) and (5), respectively.

Sensor Failure and Missing Data

When the Diagnoser is invoked due to a discrepant sensor reading, KATE assembles a list of possible suspects. The discrepant sensor itself is also placed on this list as a plausible suspect. The sensor is processed just as any other suspect object in the system, and if its innocence cannot be proven, it is retained as a potential cause for its own discrepant reading.

Sensors can, in fact, be easier to diagnose than other system objects. Validation of a sensor can, in single fault environments, become a quite trivial problem. If more than one sensor is discrepant, these sensors automatically clear one another from suspicion because no one sensor can explain another sensor's discrepant reading. The fact that KATE can diagnose a sensor failure just as any other system object and, in many situations, represent an especially easy case, distinguishes KATE's model-driven approach from conventional software and traditional rule-based techniques (3).

The ability to diagnose sensors has been of particular benefit since, historically, Shuttle launches have been threatened more by instrumentation problems than by actual system failures, and unfortunately, there did not exist a fast, reliable way to distinguish the two. The original goal of the KATE project was to provide a means to determine whether a system failure indicator in fact resulted from an actual hardware failure that might be critical to continued system operation or was merely a result of a defective instrument.

Another advantage that a model-based approach offers over a conventional approach is that missing or degraded data due to a sensor failure is easily accommodated. Whereas most conventional approaches require pre-encoded actions for each combination of sensed data being present or absent, KATE does not require any special software to handle the many combinations of missing/available data. KATE simply refrains from using any sensor data that is no longer available. Sensor data is used by the Diagnoser in two ways: (1) it is compared to expectations generated by the model during normal operation to detect faulty behavior and (2) it is compared to expectations generated from fault hypotheses to confirm or disconfirm them. If the sensor data is no longer present, it is simply not compared with the expectations generated by the model. The diagnoses that follow may result in more suspects being retained than what would have been retained had the sensor information been available, but no suspect will wrongfully be cleared. In some sense, the missing sensor data is treated the same as if the sensor data had not existed in the first place (6).

Control and Redundancy Management

Conventional control techniques require that an engineer, who is familiar with the design of a specific system and is knowledgeable of the kind of devices that make up that system and how they may fail, hard-code the functionality and connectivity of the system. Programs must be written to specify which commands need to be issued to control a specific device and which measurements need to be checked to ensure the desired goal. The programmer obviously must try to foresee all possible failures that may affect control of that device and write subroutines to handle each case.

In KATE, the system's functionality and connectivity is represented in the knowledge base and is easily and rapidly accessed by the systems thereof. This allows for the capability of low level control, based on a concept somewhat similar to that used to diagnose the cause of a failure in an engineering system. The Diagnoser infers which failed component(s) would cause a discrepant condition, whereas the Control System infers the required state of all controlling components and commands that would result in a newly desired state. This conceptual relationship resulted in the natural progression from LES to the KATE project. Using this technique, KATE is capable of

accepting a high level, desired system goal from the user, consisting of an object and a value to be obtained by the object. For instance, the user may want to open or close a discrete valve, change the state of a relay, open or close an analog valve by a certain percentage, or cause a measurement or indicator to read a certain value. KATE then determines and issues the appropriate low level commands necessary to accomplish the desired task. Furthermore, redundancy is automatically activated if required by a prior component failure.

The Control System performs the task by breaking the problem or goal down into subgoals and their subgoals and so on, recursively, until controllable objects (i.e., commands) are reached. A set of options for achieving the desired state of the object is the result of this process. For further details and examples, refer to (7). The Control System dynamically creates, selects and executes the control code. Since KATE is continually monitoring the system, checks will automatically be made to ensure successful completion of the task.

In addition, the user has the option of issuing powerful, high level commands, such that certain states of the system or specific devices are "maintained" or "controlled" by KATE. In the case that a required state of an object is affected by a failure, the Control System will automatically be invoked to search for an alternate means to maintain or control back to that requirement, and issue the necessary commands to do so.

Control Advisories

The main focus of KATE is the modeling and control of tanking systems in a launch environment. In light however, of the potential of costly damage to ground and flight hardware systems and the hazards surrounding the use of cryogenic fuel in tanking operations at the Kennedy Space Center, all involved - from KATE's software system designers to launch system operators - are understandably concerned with using the KATE prototype for on-station control until the prototype has been rigorously tested and validated. Control advisories provide an alternative to autonomous control and offer a means by which to demonstrate KATE's control capability without risk to ground or flight hardware systems. Today, for demonstration purposes, control advisories are implemented as prerequisite, reactive, or conditional control logic that is initiated upon violation of system operating criteria. Under ideal conditions, initiating the Control Advisor would be as simple as making a menu selection to instruct the KATE shell to perform advisories rather than issue commands to control hardware components. Once selected, the Control Advisor would provide recommendations on control options to the operator based on operating procedures, system constraints, and knowledge of high-level goals for control of a system or subsystem. (8)

KATE TOOLS AND UTILITIES

Generic utilities, as well as a few application specific utilities, have been made available in the KATE system. Brief descriptions of the most commonly used tools and utilities follow.

AUTONOMOUS CAMERA CONTROL - Provides visual confirmation of component status in response to an operator request or diagnosis of a component failure.

CONSISTENCY CHECKER - A knowledge base verification tool used in conjunction with the Tree Display (described below). Examines selected knowledge base slots for accuracy of information and reports any discrepancies found.

EXPLANATION FACILITY - Provides an explanation of the rationale used by KATE during diagnosis to indict or vindicate a component as a failure suspect.

KATE REMOTE DATA TRANSFER UTILITY - Allows KATE to simulate real-time data acquisition using archived data.

PLOT - Provides realtime line plots of component data for commands and measurements over time. Historical data plots may also be generated from archived data.

PROCEDURE READER - Provides the capability to enter high level goals in the form of procedural steps. KATE then executes each procedure by controlling the actual hardware or KATE's simulation of the hardware system.

SCHEMATIC DISPLAY - An interactive display of system schematics taken from actual engineering documents. Reports current status of any component depicted on the page.

SIMULATOR - Allows for simulated operation of a hardware system using a software model of the system.

SINGLE POINT FAILURE ANALYSIS - Analyzes the model and detects the "weak" points of a system, i.e. any component whose failure could cause loss of effective control of the system.

TREE DISPLAY - Provides a graphical view of the system architecture from the viewpoint of a user selected component. Displays the component and its connections to other components using the relational concepts of "upstream" (left side of the tree) and "downstream" (right side of the tree).

CURRENT APPLICATIONS

Currently, the Artificial Intelligence Section is focusing on the following two major implementations of KATE: Autonomous Launch Operations and Shuttle Liquid Oxygen Prototype.

Autonomous Launch Operations (ALO)

KATE is being developed under the U.S. Air Force's Advanced Launch Systems (ALS) Advanced Development Program (ADP) as a project entitled Autonomous Launch Operations (ALO). The objective of ALO is to demonstrate an autonomous launch control software system that performs real-time monitoring, fault detection, diagnosis and control from high-level operations requirements. It is part of an effort to reduce overall launch operations costs, by significantly decreasing process control software development and maintenance costs and by greatly reducing launch crew size, human error and fault recovery time.

In order to demonstrate the above stated goals, two hardware models have been constructed as targets for prototyping the KATE software that is being extended in order to achieve the ALO objectives: A Water (H2O) Tanking System model and an Liquid Nitrogen (LN2) Tanking System model. In addition, the existing Environmental Control System (ECS) model hardware is expected to be integrated with the tanking systems, creating a multiple-subsystem testbed, such that all three models can be used together to demonstrate a more launch-realistic environment with KATE software systems handling a variety of launch support subsystem models.

As a first step in demonstrating the KATE software for the ALO project, the H2O Tanking System model is being used for real-time monitoring, fault detection, diagnosis and control of fluid tanking systems. As part of the H2O Tanking System demonstration objectives, KATE performs the following tanking sequences and operations through its control capabilities: Ullage pressurization, Transfer line chilldown (simulated), Main Engine chilldown (simulated), Slow fill, Fast fill, Topping, Replenish, Fill circuit drain and vehicle pressurization, Engine firing (simulated) and Drainback. During these H2O Tanking System operations/sequences KATE's ability to perform numerous tasks is demonstrated. Phase 2 demonstrations are currently being performed against this H2O Tanking System hardware. As a future step, the KATE developers will be greatly challenged by the modeling issues and the required KATE software enhancements surrounding the controlling and health monitoring of cryogenic fluid systems when they move to the LN2 Tanking System model.

Shuttle Liquid Oxygen (LOX) Prototype

During a launch countdown, launch support personnel must not only monitor the health of the system and be ready to perform troubleshooting in a tense, time-critical situation, but they must also be ever aware of system constraints, operating procedures, and operating criteria. To aid in this process KATE is currently being applied to the Shuttle Liquid OXygen (LOX) loading system. KATE-LOX is funded by the Office of Aeronautics and Space Technology. The ongoing objectives of this project are to incorporate technological advances in control, monitoring and diagnostics techniques to increase productivity and reduce operator error, as well as lower software development and maintenance cost in the shuttle ground operations environment. (9)

Since April 1990, KATE-LOX has monitored fueling operations during 18 shuttle launch countdowns in an offline mode. Although the prototype is still under development, KATE-LOX has experienced success on several occasions by accurately diagnosing failures of LOX equipment during live tankings. When completed, the KATE-LOX prototype will perform monitoring and diagnosis of the Orbiter's external tank fueling operations, and will

provide advice to operators on control options.

The LOX Expert System (LES) mentioned previously, also modeled the LOX system but was implemented using a simplified model and a simplified concept of flow. KATE-LOX employs more sophisticated control and diagnostic algorithms as well as a more complex model of the LOX hardware system that more accurately represents the effects and constraints involved in the flow of a cryogen such as liquid oxygen. Also new utilities have also been designed to reflect the needs of firing room operators.

Model validation and system testing are being accomplished by exercising the system against the Shuttle Ground Operations Simulator model of the Liquid Oxygen Loading System, online monitoring during live launch countdown loading operations and simulated loadings using the KATE Remote Data Transfer Utility.

The LOX system contains approximately 515 LOX specific data points. Two-thirds of these sensor points have been modeled, the majority or which occur at the instance level. The LOX knowledge base currently contains 1750 frames and is expected the reach 2500 by knowledge base completion. (10)

SUMMARY AND FUTURE WORK

The development of KATE provides a means to avoid the creation and maintenance of large hard-coded programs to control and diagnose engineering system domains. Instead, knowledge bases are developed that describe the connections between internal application system components and how the outputs of each component depend upon its inputs (i.e., control relationships). Using this domain-specific model, KATE has the capability to intelligently control, monitor and diagnose faults for the particular application. The model produces control operations and expected values for the hardware system's measurements. Constraint checking is performed whenever a command or sensor value has changed. Upon detection of a discrepant sensor reading, the diagnoser is invoked. The model is additionally used to test diagnostic hypotheses generated as explanations for observed failures. Symbolic inversion of the dependency of a measurement upon each suspect component is used to calculate a hypothetical value for the suspect that could explain the discrepancy. Various consistency criteria are then used in an effort to eliminate all but one of the suspects - the culprit. This same inversion process is used for controlling objects by calculating input value(s) for an object which will result in a desired output. In addition, KATE uses both its control and diagnosis capabilities in performing redundancy management when a request to maintain a high level system goal is disrupted by a system failure. The model is automatically updated as the engineering system is manipulated or degrades. All diagnostic and control decisions are made in real-time, taking into account failed objects, objects which are being maintained, and those objects which are already at their desired states. Furthermore, modeling a system in terms of its structure and function, allows for the diagnosis of sensor failures similar to that of other system components. This approach also allows for easy accommodation of missing sensor information, and KATE can continue monitoring, diagnosis and control in the presence of a sensor failure. The result is increased machine intelligence in the area of reasoning about a system's health and controlling the state of a hardware system.

Uses for this type of system at the Kennedy Space Center include checkout of ground, payload, and launch support equipment, launch team training, and simulation of ground support and space station flight hardware systems for software component checkout.

Significant work remains to be done on improving the complex and time-consuming process of model building. Several avenues for the development of a graphical knowledge base editor and an automatic knowledge generation tool are being pursued to alleviate this problem. Should automatic knowledge generation reach fruition, a significant portion of KATE 's knowledge base could be generated automatically from Computer Aided Drawing system schematics and design notes.

Technical matters, such as modeling issues - particularly those relating to pseudo object detection and knowledge representation, improving diagnostic capability, and developing testing and validation methodologies continue to provide challenges. Increasing processing speed and preparing for integration with KSC launch processing systems are also concerns. To this end, porting to a conventional language and delivery platform is actively being pursued.

The development of KATE and its associated concepts are ongoing. With each new application of KATE, software enhancements are made to enable KATE to become more generic and encompassing in its ability to handle a wider variety of and more complex engineering systems.

ACKNOWLEDGEMENTS

The authors would like to recognize NASA project engineer, Tim O'Brien, of the KSC Artificial Intelligence Lab, for his assistance in the preparation of this paper. We further acknowledge the Boeing Aerospace Operations, Model-Based Systems Group for their contributions in the design and development of KATE.

REFERENCES

(1) Internal Document, "KATE Knowledge Base Generation Guide" (September 1990).

(2) R. Davis, "Diagnostic Reasoning Based on Structure and Behavior", D. G. Bobrow ed., Qualitative Reasoning about Physical Systems, MIT Press, Cambridge, MA, 1985.

(3) E. A. Scarl, J. R. Jamieson and E. New, "Model-based Reasoning for Diagnosis and Control", Proceedings of the First Florida Artificial Intelligence Research Symposium (FLAIRS-88), Orlando, FL (May 1988).

(4) E. A. Scarl, J. R. Jamieson and C. I. Delaune, "Diagnosis and Sensor Validation through Knowledge of Structure and Function", IEEE - Transactions on Systems, Man and Cybernetics SMC-17 (3), pp. 360-368 (May/June 1987).

(5) S. Thomas, "Symbolic Inversion of Control Relationships in Model-Based Expert Systems", Final Report - NASA Research Grant NAG10-0045 (December 1988).

(6) C. L. Belton-Parrish and S. Enand, "KATE - A Model-based Diagnostic and Control Shell", Intelligent Diagnostic Systems, Eds. K. F. Martin, J. H. Williams and D. T. Pham, IFS/Springer-Verlag, to be published in Fall 1992.

(7) E. New, "Knowledge-Based Control and Redundancy Management Techniques Used in NASA's KATE Project", Proceedings of Southcon/87, Orlando, FL (March 1987).

(8) T. Gould, "KATE Video Script", KATE 20 Minute Video (April 1991).

(9) C. L. Belton and B. L. Brown, "Knowledge-Based Autonomous Test Engineer (KATE) - A Model-Based Control and Diagnostic Shell", Research and Technology, 1990 Annual Report, NASA Technical Memorandum 103811.

(10) B. L. Brown, "KATE-LOX Narrative", Code RC Center Management Review (August 1991).

N92-22686

ADVANCED COMPUTED TOMOGRAPHY INSPECTION SYSTEM (ACTIS): AN OVERVIEW OF THE TECHNOLOGY AND ITS APPLICATION

Lisa H. Hediger
National Aeronautics and Space Administration
Nondestructive Evaluation Branch
Mail Stop EH13
George C. Marshall Space Flight Center
Huntsville, AL 35812
Phone: (205) 544-2544

ABSTRACT:

The Advanced Computed Tomography Inspection System (ACTIS) was developed by the Marshall Space Flight Center (MSFC) to support in-house solid propulsion test programs. ACTIS represents a significant advance in state-of-the-art inspection systems. Its flexibility and superior technical performance have made ACTIS very popular, both within and outside the aerospace community. Through Technology Utilization efforts, ACTIS has been applied to inspection problems in commercial aerospace, lumber, automotive, and nuclear waste disposal industries. ACTIS has even been used to inspect items of historical interest.

ACTIS has consistently produced valuable results, providing information which was unattainable through conventional inspection methods. Although many successes have already been demonstrated, the full potential of ACTIS has not yet been realized. It is currently being applied in the commercial aerospace industry by Boeing Aerospace Company. Smaller systems, based on ACTIS technology are becoming increasingly available. This technology has much to offer the small businesses and industry, especially in identifying design and process problems early in the product development cycle to prevent defects. Several options are available to businesses interested in pursuing this technology.

BACKGROUND:

Computed tomography (CT) is a nondestructive inspection method which provides a cross-sectional view of an object using penetrating x-rays. CT images are extremely useful for inspection of complex objects because they provide a visual representation of defects without the scatter and superpositioning problems commonly encountered in conventional x-ray inspection.

In CT, the x-ray beam passes through the object under test and is absorbed to various degrees as it travels through. The nonabsorbed x-rays strike an array of electronic detectors. Each individual detector in the array is discrete. That is, each detector collects the transmitted x-rays along a sharply defined line of sight between it and the x-ray source. As the part location is systematically changed and the x-ray measurements are repeated at each new location, a data set consisting of several million individual line x-rays is compiled. With the help of a fast array-processing computer, this collection of line x-rays from different orientations is mathematically combined to reconstruct an image of the object cross-section at a particular height.

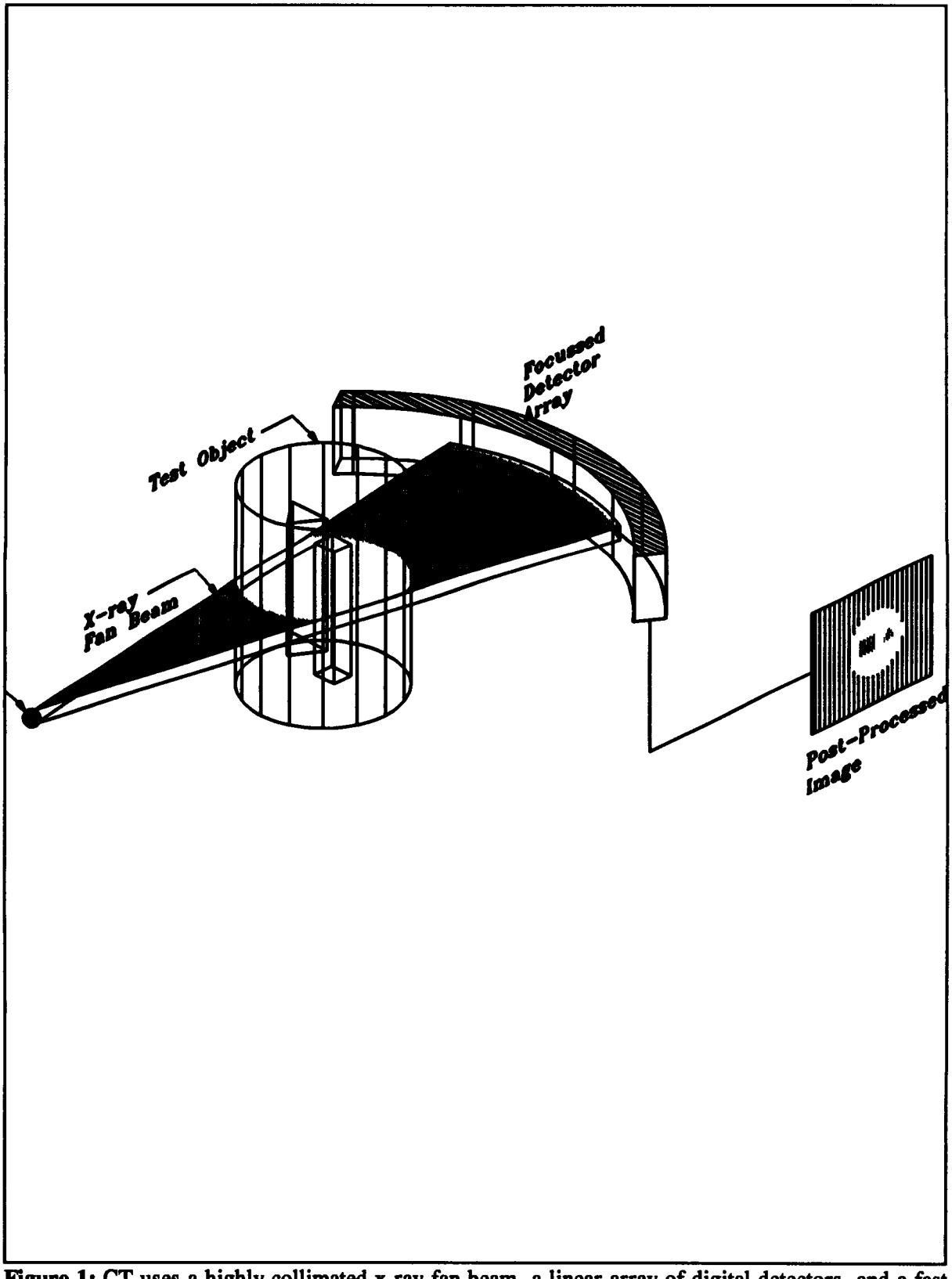


Figure 1: CT uses a highly collimated x-ray fan beam, a linear array of digital detectors, and a fast array-processing computer to produce a cross-sectional image of the object.

Computed Tomography (CT) is becoming popular with researchers in the areas of composite materials and liquid and solid propulsion. CT provides a great deal of information in a digital format, which can be stored on a compact media, recalled quickly, and digitally enhanced to highlight details of interest.

In contrast to other nondestructive evaluation techniques, CT places fewer restrictions on component configuration and material selection. Unfortunately, the design of a CT system is usually optimized for a known type of material or component size, to allow the manufacturer to produce a system which will perform satisfactorily for the required application at minimal cost.

Eight years ago, Marshall Space Flight Center (MSFC) became increasingly involved with composite materials and solid propulsion research. CT was attractive because it provided meaningful quantitative data. In the early days, hospital CT systems were used to evaluate small samples. However, the need to evaluate subscale test and small flight items soon developed. A complete needs and requirements analysis revealed a need to evaluate components ranging from four inches to four feet in diameter and materials ranging from steel to rubber. Available CT systems were not flexible enough to meet these requirements, so the Advanced Computed Tomography Inspection System (ACTIS) was developed as a cooperative effort between MSFC and Bio-Imaging Research (BIR).

Most x-ray CT systems consist of a single x-ray source, a linear array of approximately 125 detectors, an array-processing computer, and peripheral devices. ACTIS system consists of four high-energy x-ray producing devices, an integrated array of 625 solid-state x-ray detectors, a 162 channel data acquisition system, mechanical gantry with 13 degrees of freedom and its control computer, a fast parallel image-processing computer, an integrating operator console, and peripheral devices.

Typically, the performance of a CT system is determined by the following characteristics:

- Suitability for application
- Spatial resolution
- Contrast sensitivity
- Scan speed
- Ease of use/maintenance

Each of these performance characteristics affects cost. ACTIS provides superb technical performance at a competitive price. ACTIS is the most flexible CT system available. Its performance, in terms of spatial resolution, scan speed, and contrast sensitivity are exceptional. Its flexibility and technological merit have made ACTIS one of the most heavily-loaded industrial CT systems in the world.

NASA applies ACTIS to investigate the behavior of new and exotic materials in the ballistic environment. Test motors, incorporating these new materials, are assembled and test-fired at MSFC. Pre- and post-fire CT analysis reveals how the structure of the material is altered by ballistic conditions. ACTIS is sensitive to changes in material density of about 0.05 percent, so it is sensitive to even slight changes in material structure. Typical scan times per slice range from seven minutes to thirty minutes, depending on the object size and opacity, and the required contrast sensitivity, and spatial resolution.

Visualizing defect location is sometimes a problem with CT, since it is difficult for most laymen to think in terms of slices. ACTIS employs Multi-Planar Reconstruction (MPR) as a visualization tool for CT-located defects. In MPR, individual CT slices are digitally stacked and displayed from each of three perpendicular viewing angles. Visualizing defect location with MPR is typically quite easy for

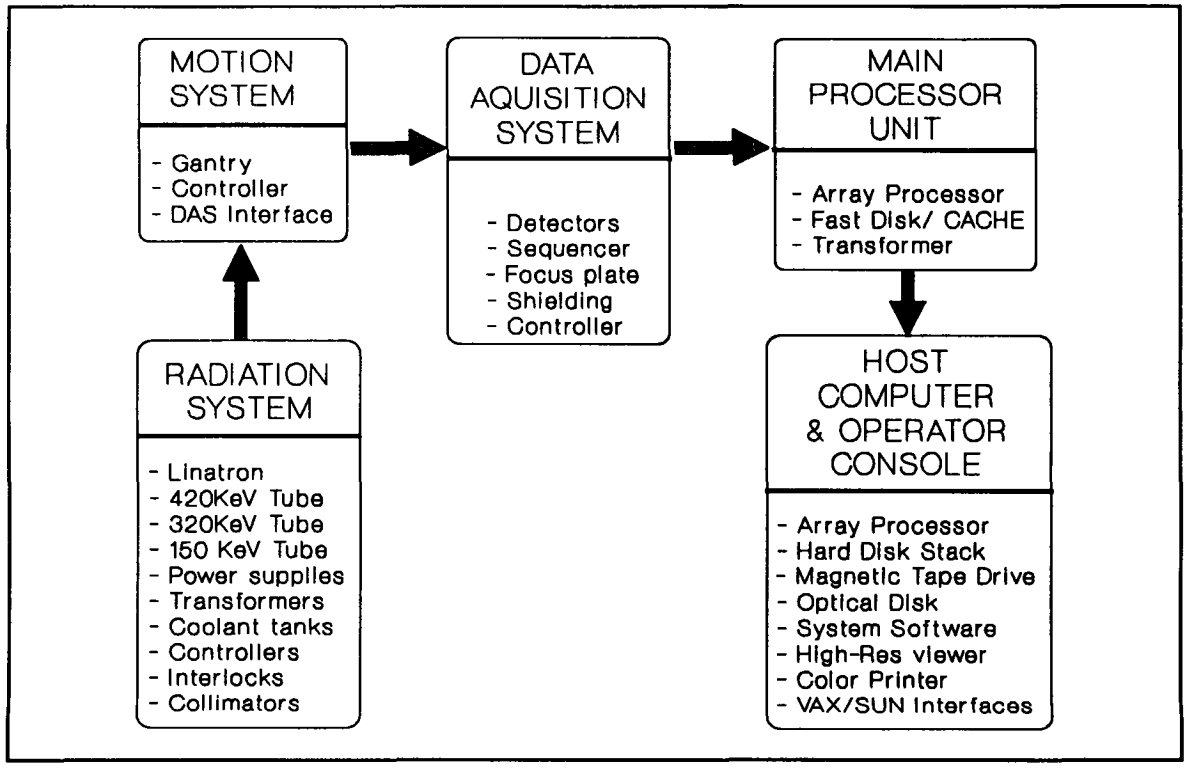


Figure 2: Simple Block Diagram of ACTIS

technical personnel, who are used to dealing with engineering drawings.

A number of special image-processing features are available with ACTIS. These features are available from the main control console and from a remote imaging workstation, which is interfaced to the ACTIS host computer. Special features include contrast adjustment, statistical image analysis features, relative density plots along a line through the image, zoom magnification, and shaded surface rendering based on a stacked set of CT images.

Although the cost of ACTIS is competitive with most large industrial CT scanners, CT in general is an expensive method of inspection due to long scan times, the need for semi-skilled operators, and the acquisition / maintenance cost of the equipment. In fact, CT is so expensive that in most cases, it makes little financial sense to use it for routine inspection. However, during the early stages of new product development, it provides information on design and process weaknesses which can save money in the long run. Recently, ACTIS significantly affected the development of an improved version of the Space Shuttle Main Engine (SSME) Turbopumps by spotting anomalies in the casting process during the early development phase of the program. These defects (slight imperfections in the internal walls) were impossible to detect with conventional inspection methods. Without ACTIS, the program may not have ever known they existed. Further, the program discovered rework of these defects was difficult and costly. As a result of continued CT analysis, minor changes made to the casting process eliminated these costly defects at the source.

In addition to NASA programs, ACTIS has benefitted many military programs. It has been used to inspect thermal batteries for Air Force systems, artillery storage canisters for Army depots, and housings for Navy cruise missiles. Most of these agencies have special-purpose CT systems of their own.

Figure 3: ACTIS Mechanical Gantry with 13 degrees of freedom accommodates objects ranging in diameter from 4" to 4' and weighing up to 2000 pounds.

One of the most unique applications of ACTIS was the inspection of a 100-year-old time capsule for *National Geographic* magazine. The time capsule, commemorating the centennial of the inauguration of George Washington, had been sealed in 1889. Before the box was opened, NASA had the opportunity to scan the contents with ACTIS. Originally, ACTIS was intended to identify how the box was assembled, and to offer suggestions on how to open it without damaging the contents. ACTIS far exceeded our expectations, providing not only information about the construction of the box (wood sandwiched between two layers of tin), but details pictures of the contents. When scanning was complete, NASA had tentatively identified five medallions, the letters on a calendar, and several books. Our findings were later verified through image enhancement performed at Technical and Analytic Sciences Corporation.

ACTIS was used to perform feasibility studies for applications of CT in the lumber industry. These studies revealed that CT provides useful information to the lumber mills, allowing them to grade lumber, optimize cutting plans to maximize yield, and cut costs by identifying scrap prior to finishing.

ACTIS has been used by the U. S. Department of Energy to inspect barrels of stored nuclear waste. The system can identify free liquids trapped within the waste and structural flaws in the barrels themselves, two significant hazards in nuclear waste storage. Using ACTIS, scientists are able to identify barrels which present an unacceptable risk, and take preventative action. ACTIS was also used by the DOE to identify the contents of waste barrels. To test these capabilities of ACTIS, the DOE provided a 55-gallon drum of simulated nuclear waste. In the barrel were several common objects: rubber o-rings, a pen, a filament, a comb, a lead-lined rubber glove, and others. ACTIS operated blind. That is, the operators were not told anything at all about the contents of the barrel. The tests produced some impressive results. ACTIS was able to image the face on a dime contained in the 55-gallon drum of waste.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Various automotive manufacturers have explored the use of ACTIS in designing new products. The system has been used to inspect prototype steering wheels, engine blocks, gear boxes, and other structures for automotive applications. ACTIS has become an integral part of prototype development for some of these companies. American automotive manufacturers use CT data to evaluate performance of new designs and processes early in the development cycle to reduce the potential for defects in routine production.

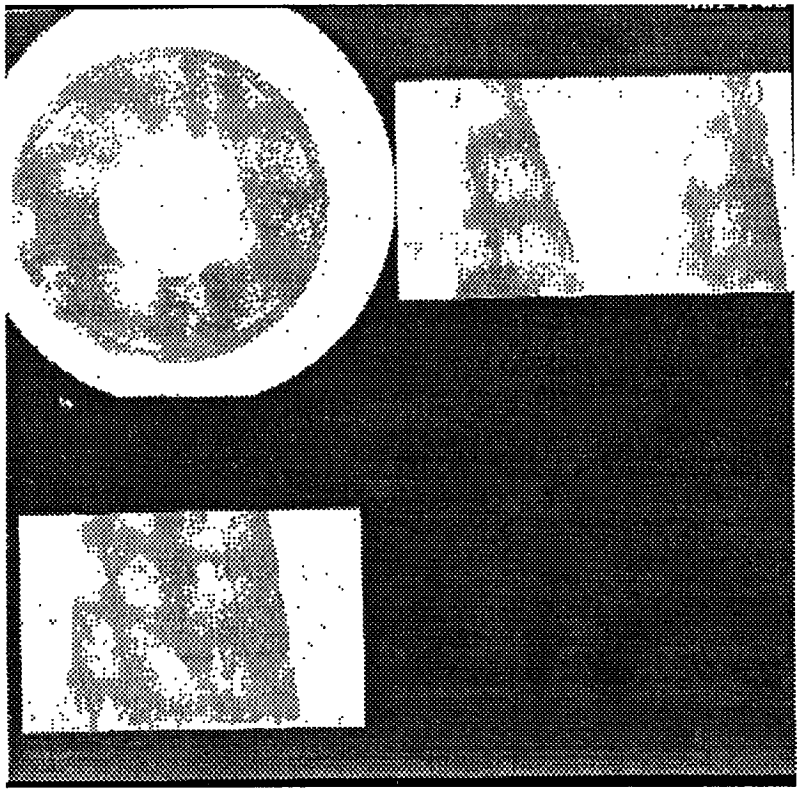


Figure 4: Multi-planar Reconstruction of a Rocket Nozzle Exit Cone clearly represents the location of defects.

ACTIS systems are now sold as off-the-shelf items. Unfortunately, these systems are too expensive for most small businesses to afford. However, there are several alternatives for small businesses and industries interested in applying ACTIS technology. First, various CT system manufacturers can tailor a system for a specific need at substantially reduced cost. NASA recently purchased an ACTIS-II system, a much smaller and somewhat modified version of ACTIS, at about 5% of the cost of the original system. Interested businesses can often lease large industrial scanners for a few days at a time. Several large industrial scanners are available for lease. If the program is well-planned, this can be a cost-effective way to use ACTIS technology. Finally, if the object under test is small and portable, small businesses can buy CT services at a medical radiology laboratory. NASA did all its early work with CT using a local hospital after-hours. For certain materials, the medical systems actually perform superior to industrial ones.

CONCLUSIONS:

ACTIS is a versatile inspection tool which has proven useful in addressing many of the problems associated with inspecting complex objects, not only in the government aerospace industry, but other industries as well. NASA has applied this technology to the problems of material science for propulsion systems. A variety of other industries have benefitted from ACTIS technology. ACTIS results are superior to conventional inspection techniques in flexibility, contrast sensitivity, spatial resolution, and visualization. CT inspection typically imposes fewer restrictions on the geometric and other physical features of the object under test. Although CT systems are quite expensive, the results they provide are useful, especially early in the development of new products, by allowing engineers to detect basic design flaws and process anomalies early enough to eliminate them. Small business can take advantage of ACTIS technology in a number of ways: buying a smaller system, leasing and industrial system, or buying time

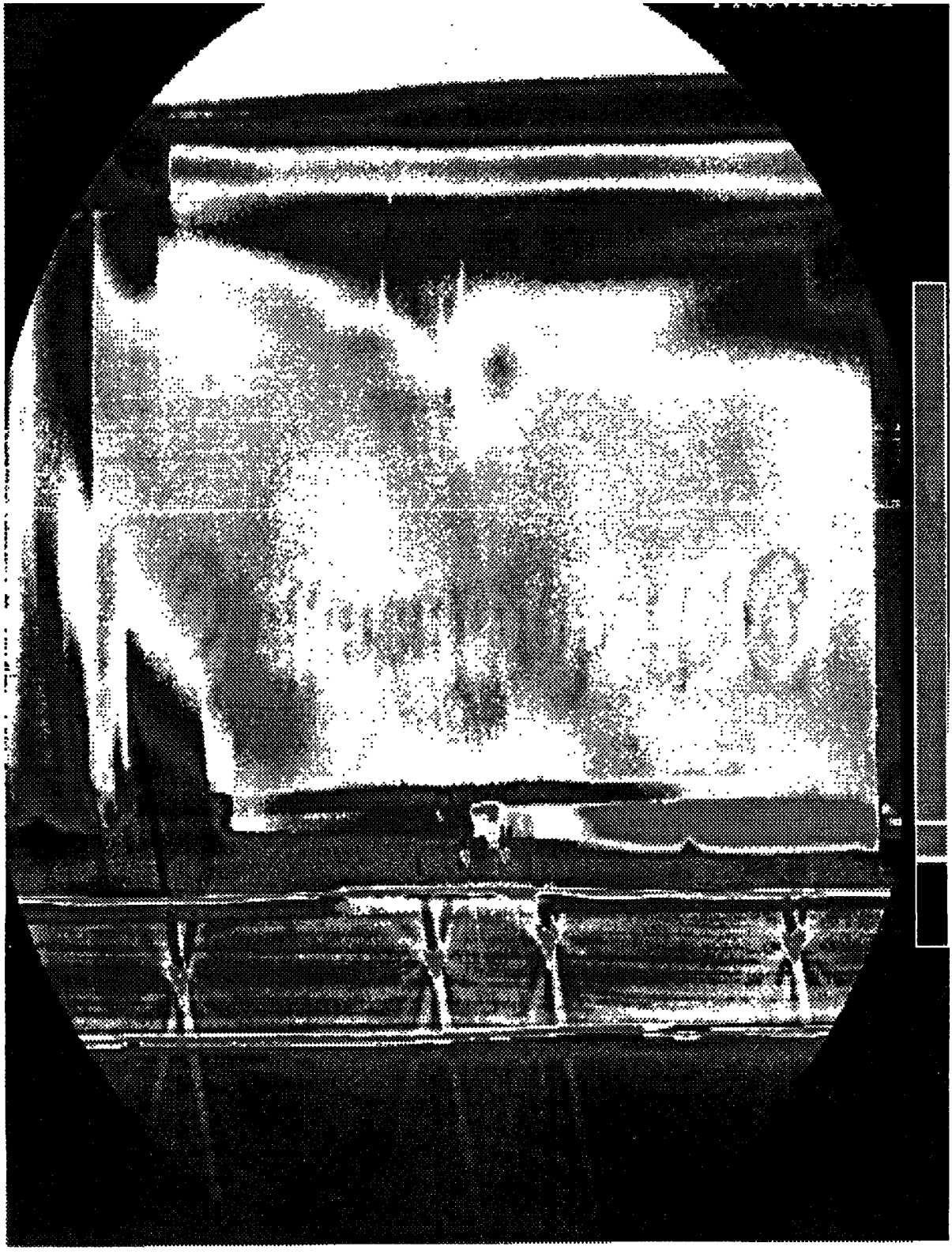


Figure 5: Unenhanced image from time capsule. Note the faint letters spelling out "Stettiner-Lambert & Co." These letters turned out to be written on a calendar inside the capsule.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

on a medical system. Although ACTIS is not the answer to all inspection problems, it provides small business and industry with an excellent tool for building quality into its products.

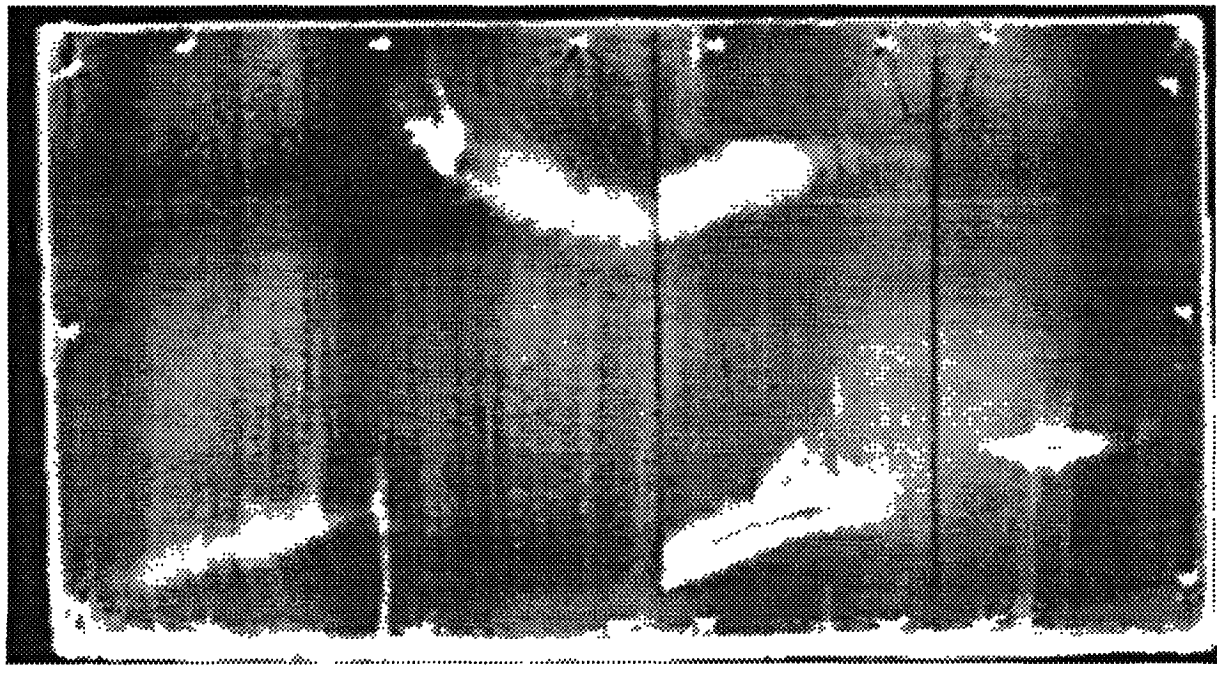


Figure 6: Image of time capsule lid. Utility of ACTIS for lumber industry was discovered here. Note the grain of the wood is visible, as well as separations between planks, and nail locations.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

N92-22687

HIGH RESOLUTION ULTRASONIC SPECTROSCOPY SYSTEM FOR NONDESTRUCTIVE EVALUATION - SBIR PHASE III

Dr. C. H. Chen
Information Research Laboratory, Inc.
415 Bradford Place
North Dartmouth, MA 02747-3819

ABSTRACT

With increased demand for high resolution ultrasonic evaluation, computer-based systems or work stations become essential. In this project, the ultrasonic spectroscopy method of nondestructive evaluation (NDE) has been employed to develop a high resolution ultrasonic inspection system supported by modern signal processing, pattern recognition, and neural network technologies. The basic system which has been completed consists of a 386/20 MHz PC (IBM AT compatible), a pulser/receiver, a digital oscilloscope with serial and parallel communications to the computer, an immersion tank with motor control of X-Y axis movement, and the supporting software package, IUNDE for interactive ultrasonic evaluation. Although the hardware components are commercially available, the software development is entirely original. By integrating signal processing , pattern recognition, maximum entropy spectral analysis, and artificial neural network functions into the system, many NDE tasks can be performed. The high resolution graphics capability provides visualization of complex NDE problems. The Phase III efforts of this SBIR project involve intensive marketing of the software package and collaborative work with industrial sectors.

INTRODUCTION

This paper presents a low cost high resolution ultrasonic inspection system that provides extensive software suppoort to many nondestructive tesing tasks. Ultrasonic method of inspection is least costly but its capability can be limited because of hardware constraints. In any test and evaluation, a good set of measurements is most desirable. In practice the information available from the measurements is limited. Digital processing of the measurements offers the possibility to extract a lot more information from the data with the use of modern signal processing and related techniques. The design approach of the system developed is to rely entirely on the digital signal processing, pattern recognition and neural networks for defect characterization, detection and classification. Fig. 1 shows a block diagram of the system which consists of a Panametrics 5052UA pulser/receiver, a LeCroy 9400 digital oscilloscope, a National Instruments GPIB board, a Dell System 310 computer with math co-processor, a Testech immersion tank with X-Y axis motor control unit, a line printer and a plotter. Fig. 2 is a photo showing the pulse echoes from a small lead ball in the immersion tank. The lower figure is a smoothed digital signal of the upper figure taken from the digital scope. There are many other commercial components which may serve the same NDE purpose. The idea is that only commercially available parts are used. The innovative element of the system is the extensive signal processing software support which will be described in the next section.

The term "ultrasonic spectroscopy" was probably first usad by Otto R. Gericke of US Army Materials Technology Laboratory in early 60's, who found experimentally a strong correlation between the frequency spectrum of the pulse echoes and the geometries of the hidden defects. Although there were a number of subsequent efforts in the NDE community to use ultrasonic spectroscopy in other material testing problems, the success has been limited because of the lack of quantitative relationships between the defects and the frequency spectrum. However, the spectral domain methods are fundamental to modern signal processing, and by expanding ultrasonic spectroscopy to a larger class of spectrum based signal processing techniques, much more information about the defects can be extracted from the data. The word "high resuliton" refers to the cability to examine fine details of the signal and thus the defects. Such high resolution capability can only be offered by the software without added cost to the system.

THE IUNDE VERSION 2.3

Currently the software package IUNDE is in Version 2.3 and contains three parts.

Part 1. The signal processing package.
Signal acquisition: acquire signal from a disk file (in ASCII format), acquire signal from STR*825 board, acquire signal from LcCroy 9400 digital scope.

Power spectra: the fast Fourier transform (FFT), the Burg's maximum entropy spectrum analysis; the chirp-z transform; correlation; bicorrelation and bispectrum.

Deconvolution by Wiener filtering, by spectral extrapolation and by least mean square error (LMSE) criterion.

Special transforms: analytic signals, Hilbert transform, power cepstrum, discrete pseudo Wigner distribution (DSWD) and wavelet transform for time-frequency analysis.

Preprocessing: mean removal, moving average, circular shift, zero-padding, and amplitude normalization.

Graphic display using GraphiC software version 5.0 for 1-D, 2-D, 3-D and multi-window displays.

Feature extraction which extracts the following features: mean and maxima, peak correlation with reference, amplitude ratio, frequency ratio, moments of the spectrum, bandwidth and frequency of peak power, fractional power distribution in 8 bands.

Automated defect classification using Nestor's NDS-100 Neural Network.

Automated defect classification using back-propagation neural network.

Part 2. The pattern analysis package

In IUNDE, two different approaches to pattern recognition are available. The first approach is by using traditional statistical pattern recognition techniques with functions including the k-mean clustering, the nearest neighbor classification, Bayesian classification, Foley-Sammon transform, multiple Fisher's linear discriminant as well as nonlinear mapping. The second approach uses the neural network as described in Part 1.

Part 3. The MESA (Maximum Entropy Spectral Analysis) package

It includes four methods of spectral estimation: Burg, modified covariance, FFT and Broyden methods, with up to 4096 data points allowed as input and up to 8000 output spectral points.

REPORT OF THE PHASE III

The results of Phase II were presented in detail in the Final Report submitted to the Materials Technology Laboratory (MTL) in December, 1990. Since then continued work has been done to improve the software package. Intensive marketing of the software package has been the major Phase III effort. Eleven US companies and five foreign companies representing four countries have installed the software package in their NDE systems. Collaborative work is performed with one company to adapt the package to NDE of composite materials. Much more effort is needed to seek for companies which are interested to fund the collaborative work.

CONCLUDING REMARKS

Technically speaking, the system with the software package has incorporated the best available signal processing, pattern recognition and neural network technologies in the ultrasonic NDE of materials. The use of modern signal processing and related techniques as reflected in the entire system design can offer the truly high resolution ultrasonic NDE capability much needed for many applications. Future NDE systems are likely to equip with many of such techniques. The rapid development in signal processing hardware (such as the DSP chips) and advances in computation capabilities will make such systems even most cost effective. There is no doubt that the trend in NDE industry is to make increased use of digital signal processing software and hardware. On the other hand, the market place has been slow to follow such trend. Also a good communication between the digital scope and the computer is not an easy problem. Other hardware designs have incorporated the the pulser/receiver and the high speed digitizer in the computer but the role of digital scope is not fully served with such arrangement. Continued software and hardware improvements are definitely needed to effectively utilize such a high resolution system in a variety of NDE tasks.

ACKNOWLEDGEMENTS

I am most grateful to Dr. Otto R. Gericke of MTL (now retired) for his guidance and encouragements throughout most of this contract (Phases I & II). I also would like to thank all members of MTL Nondestructive Evaluation Group for their encouragements and interests in this work.

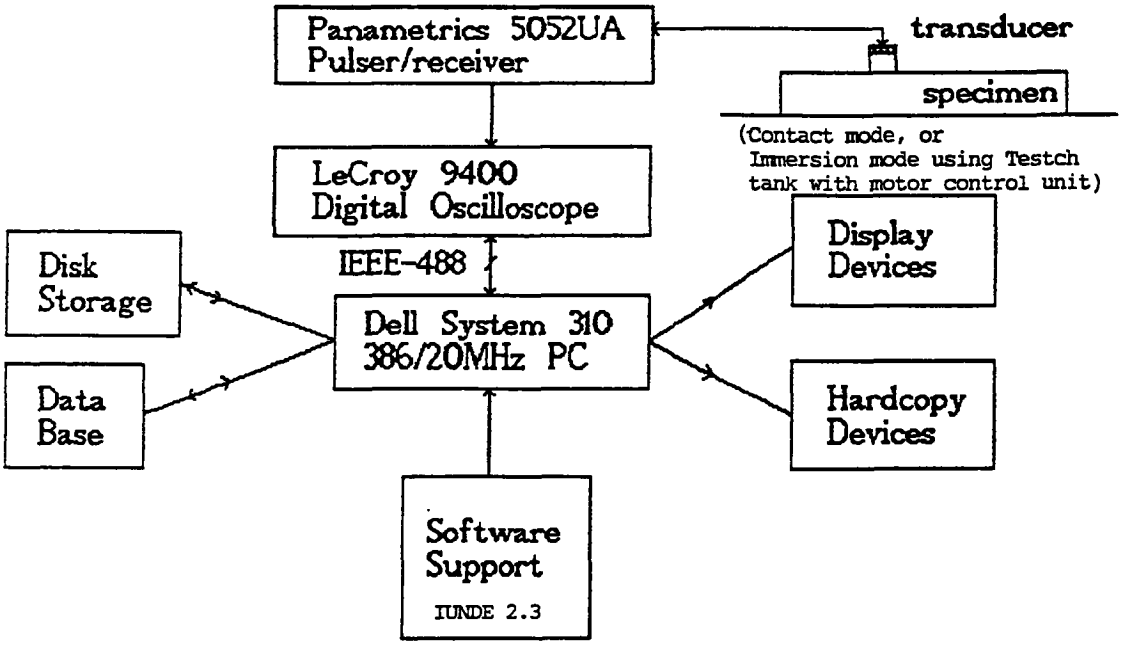


Figure 1. A block diagram of the high resolution ultrasonic spectroscopy system

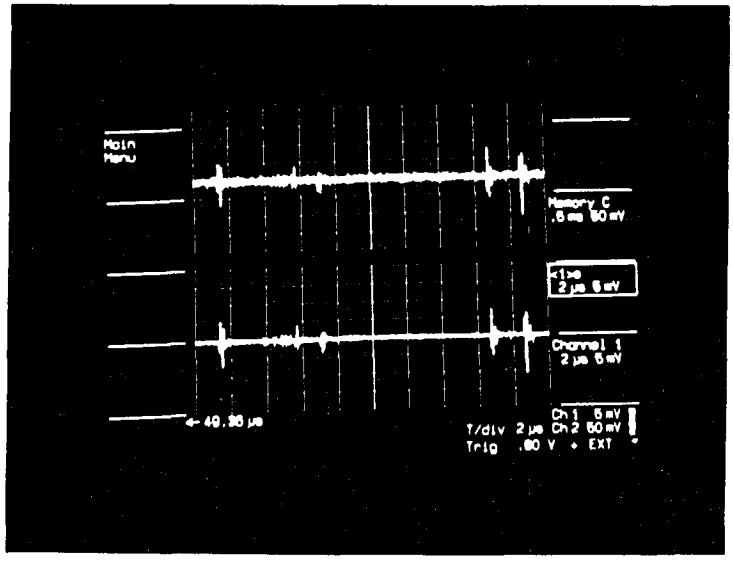


Figure 2. A typical long-window signal shown in digital scope. The upper figure shows the pulse echoes from a small lead ball in the immersion tank with a 10MHz transducer. The lower figure is the smoothed signal of the upper figure. Only two channels of signals are shown here.

ORIGINAL PAGE IS
OF POOR QUALITY

N92-22688

FORCE LIMITED VIBRATION TESTING

Terry D. Scharton
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

ABSTRACT

A new method of conducting laboratory vibration tests of spacecraft equipment has been developed to more closely simulate the vibration environment experienced when the spacecraft is launched on a rocket. The improved tests are tailored to identify equipment design and workmanship problems without inducing artificial failures that would not have occurred at launch. These new, less-destructive type of vibration tests are essential to JPL's protoflight test approach in which laboratory testing is conducted using the flight equipment, often one-of-a-kind, to save time and money. In conventional vibration tests only the input vibratory motion is specified; the feedback, or reaction force, between the test item and the vibration machine is ignored. Most test failures occur when the test item goes into resonance, and the reaction force becomes very large. It has long been recognized that the large reaction force is a test artifact which does not occur with the lightweight, flexible mounting structures characteristic of spacecraft and space vehicles. In the new vibration tests, both the motion and the force provided to the test item by the vibration machine are controlled, so that the vibration ride experienced by the test item is as in flight. Reaction force limiting has been used successfully at JPL in three vibration tests of flight instruments during the past year, and it is anticipated that the new technique will be widely employed soon at JPL and eventually throughout the aerospace industry.

INTRODUCTION

The acceleration input specified in a vibration test of aerospace hardware typically exceeds the relevant flight input for several reasons. Figure 1 illustrates the relative effects of smoothing, enveloping and adding margins to flight vibration data. The test specification tends to smooth, or average, over the frequency variations in the flight data because these variations depend on the details of the flight mounting and attached hardware. A consequence of this smoothing is the elimination of flight data interface acceleration notches which may be caused by the large reaction forces at resonances of the attached hardware. These innocuous looking notches are very important because vibration testing with a smoothed acceleration specification may result in overtesting at the resonance peaks of lightly damped test items by up to a factor of one-hundred. (The factors discussed in this example are applicable to linear amplitudes of vibration. For spectral densities, square the factors; for dB levels, take twenty times the log of the factors.) In most cases the effect of not notching the acceleration test spectrum overshadows the effects of enveloping the flight data to account for spatial, flight-to-flight, and other variations, and of applying margins to the flight data for design and test. Enveloping factors are typically in the range of two to three, which generally include the 95th percentile. Design and test margin factors range from one-and-a-quarter to two; the Jet Propulsion Laboratory (JPL) uses one-and-a-half.

Figure 2 illustrates a zoom analysis of the frequency dependence of force and accelerations at one of the notches shown in figure 1. The force at the interface between the mounting structure and the test item, the interface acceleration, and the response acceleration of the test item are plotted as functions of frequency for both flight and vibration test mounting configurations. The interface acceleration is notched in flight but smooth in a conventional vibration test, as discussed previously. The interface force increases moderately at a test item resonance in flight, as the interface acceleration notches down. The interface force increases greatly at a test item resonance in a vibration test with constant interface acceleration. The ratio of interface force to interface acceleration is the same in flight and in a test; the ratio is given by the test item apparent weight which has a peak at the test item resonance. The response acceleration of the test item has the same

frequency dependence as the interface force; in fact, the response is equal to the interface force times the test item modal mass for the resonance. The plots in figure 2 show that for flight mounting, the interface force increases to a limit value at resonance and the interface acceleration decreases from a limit value. Limiting the interface force in a vibration test (that is the force delivered to the test item by the shaker) is equivalent to limiting the test item response at resonances.

THEORY

The following exact relationship for the dual control of shaker acceleration and force has been the subject of several recent investigations [1-7]:

$$1 = A/Ao + F/Fo \quad (1)$$

where A is the acceleration of the source, Ao is the source free acceleration, F is the force between the source and load, and Fo is the source blocked force. The quantities in Eq. 1 are complex functions of frequency and it is difficult to predict or measure the associated phase angles. It may be seen from the results in [5] that the interface acceleration A and force F may both exceed the respective free Ao and blocked Fo values at the resonances of the coupled system. Therefore, the magnitudes of these quantities do not obey equation 1, but equation 1 is satisfied when phase is taken into account [5].

A more useful relation for vibration testing is the extremal dual control equation [1,4]:

$$|A| < A1 \text{ and } |F| < F1 \quad (2)$$

where A1 and F1 are limit magnitude values appropriate to a vibration test specification. To be useful for extremal control of a shaker, Eq. 2 must be applicable at all frequencies and particularly at the coupled system resonances where as shown in [5] the free acceleration and blocked force are not adequate limits. It is essential that the acceleration limit A1 and force limit F1 in Eq. 2 be chosen so that the respective coupled system values are enveloped.

FORCE SPECIFICATIONS

There are a number of approaches available for defining the force limit [6,7] in Eq. 2. Herein, the source impedance method [2,4 & 7] is described. The acceleration limit in Eq. 2 is that normally used to specify a conventional hard-base-drive vibration test, i.e. the envelope of the interface acceleration between the source and load. In a random vibration test the acceleration specification is the spectral density As in G**2/Hz. In the source impedance method, the corresponding force spectral density limit is calculated as the product of the acceleration specification As and the square of the source effective mass Ms in lb mass. If the measured force in the test set-up includes the force used to move a test fixture of mass Mf, the appropriate total force spectrum Fs in lb**2/Hz is:

$$Fs = As * [Ms**2 + Mf**2] \quad (3)$$

where it has been assumed that the fixture force and load force are 90 degrees out of phase at load resonance (see Fig. 13 of [4]).

The force spectrum calculated using the source impedance method will in general differ from that calculated more accurately using the method in [6] or the two-degree-of-freedom method in [7]. The correct way to calculate the force is to multiply the load mass M2 times the load acceleration A2, whereas with the source impedance method the source mass M1 is multiplied times the source acceleration A1. The factor needed to correct the source impedance force spectrum is the product of the ratio of mean-square (or peak) load to source responses times the ratio of squared load to source masses. For M2/M1 <0.4, the source

impedance method yields forces larger than the two-degree-of-freedom method and is therefore conservative for qualification purposes (see Fig. 3 in [7]). For M2/M1 > 0.4, the source impedance method becomes unconservative for qualification, and the more exact calculation methods of [6 and 7] must be used.

An extremal control approach, similar to that described in [4], can be used to automatically implement force limiting. In the extremal approach, the shaker control system compares several measurement channels with appropriate reference spectra and adjusts the shaker drive until one channel is equal to the reference and the other channels are equal to or less than their references. When one channel is force and the other is acceleration, the extremal dual control approach of Eq. 2 is automatically implemented. Unfortunately, most shaker controllers, including JPL's, currently provide for only one reference spectrum.

Figure 3 shows the flow diagram which can be used to implement extremal dual control with existing test equipment. Channels 1 and 2 are redundant control accelerometers in the shake direction and channel 3 is the force transducer signal. S1 and S2 are the control accelerometer charge amplifier sensitivities in volts/g, Sf is the force transducer charge amplifier sensitivity in volts/lb, and S3 is the pseudo-accelerometer sensitivity in volts/g input to channel 3 of the controller. The one-third octave spectrum shaping filter gain settings are calculated from S3, Sf, and the acceleration As and Force Fs specifications as shown in Fig. 3.

The effective source mass Ms can be measured as a function of frequency with a modal impact hammer, if the mounting structure of the test item is available. Alternatively, the source mass can be determined from a finite element model of the mounting structure. Data must be obtained in each of three directions at each of the mounting points. The raw data will show peaks and valleys associated with anti-resonances and resonances. The effective mass is a smooth curve through the data. Foster's theorem says that the effective mass is a decreasing function of frequency [8]. If a finite element model is used, the effective mass is the sum of the modal masses of all modes with resonant frequencies at and above the frequency of interest. The total effective mass in each direction is calculated by summing the masses at each of the attachment points.

TEST RESULTS

A vibration retest of a spacecraft flight instrument, the Mars Observer Camera (MOC) was conducted using force limiting [9]. The MOC is a principal investigator supplied instrument being built by the California Institute of Technology Division of Planetary Sciences for Arizona State University. The MOC instrument is part of the JPL Mars Observer spacecraft configuration scheduled for launch in 1992. Figure 4 is a schematic of the MOC configured for a vertical axis vibration test. The MOC consists of an f/10 reflector telescope with approximately a 14" aperture and a graphite epoxy tube approximately 30" in length. The camera also includes a wide angle lens assembly located along one side of the tube.

The test fixture consisted of a 2" thick, 18" diameter, 50 lb aluminum plate to which the MOC was attached at three mounting feet as shown in Figure 4. The input acceleration was controlled using the extreme of two control accelerometers mounted on this fixture plate. Four triaxial piezoelectric force transducers (Kistler model 9067) were sandwiched between the fixture plate and a 2", thick 20" diameter shaker base plate which was attached to the shaker head. To enable the transmission of shear forces across the transducers in the lateral tests, the force transducers were preloaded axially to 8400 lb with a 1/2" through bolt torqued to 70 ft-lb. The outputs of the four force transducers in the test axis direction were summed to provide total lateral shear or vertical force. In the horizontal axis tests, the vertical direction outputs of the two transducers further from the shaker were summed and multiplied by two to estimate the total vertical moment force. The summed transducer output was attenuated by a charge amplifier to accommodate the high sensitivity force transducers. The output of the charge amplifier was sent to a one-third octave spectrum shaping network and the shaker controller for force limiting as described in Figure 3.

The MOC vibration test was conducted in JPL's environmental test facility during the two day period January 19 and 20, 1991. After some initial difficulties setting up JPL's data acquisition system to record the force data and trouble shooting accelerometer channels, the tests went smoothly. The additional time associated with implementing the force limiting technique for the first time was saved by not having to calculate and implement manual notching. The fact that a complicated three-axis test was completed in two, admittedly long, days speaks to the efficiency of the technique. The MOC was well instrumented with accelerometers and the response data at critical locations was analyzed between runs before going to higher test levels. The MOC passed the test without any structural or performance degradation.

Figure 5 shows the measured vertical force in the minus 18 dB vertical random vibration test without and with force limiting. Also shown is the force specification calculated from Eg. 3 by taking the fixture plus mounts (squared mass) times the acceleration specification and subtracting 18 dB. The force limiting reduced the force peak at 285 Hz by about 9 dB. However, some off-resonance response below 100 Hz is also reduced. Following the -18 dB run, some adjustments were made to the force specifications by changing the one-third octave filter gain settings; specifically, the non-resonant force limiting below 100 Hz was eliminated and the amount of force limiting at the first resonance at 285 Hz was increased.

Figure 6 compares the control acceleration spectrum for the full level vertical random vibration test with the 0.2 G**2/Hz specification. The force limiting resulted in approximately a 15 dB notch at the fundamental resonance at 285 Hz. Notice that the 285 Hz notch is a gradual ramp down and a sharp step increase corresponding to the mirror image of the force excess of the specification in Fig. 5. This asymmetric notch shape is a characteristic of the force limiting approach. It is believed that this notch shape is more representative of flight than the symmetric notches typically used in manual notching to limit response acceleration to calculated limit loads. The notches at 400 and 630 Hz were put in by increasing the filter gain in these one-third octave bands after the -12 dB run in order to keep the acceleration responses at two critical locations under the limit load levels. Unfortunately, these resonances were masked by the large force required to move the 50 lb fixture, so that the notches are one-third octave wide instead of being narrower in width like the resonances.

CONCLUSIONS

Force limiting has been utilized successfully at JPL in three vibration tests of flight instruments, one of which was the MOC described herein. In each case, the test item received a softer ride than it would have in a conventional vibration test with only acceleration control. However, the author is convinced that each test was a realistic representation of the flight environment, plus some margin. Force limiting offers a rational means of eliminating the costs and schedule delays associated with both overdesign and overtesting in aerospace and automotive industries.

Wide spread application of force limiting will require more experience and flight force data to develop generic force specifications. Techniques for predicting the overturning moments in lateral axis tests and for combining the effective masses at multiple mounting points are needed. Improved force gage mounting methods are needed to alleviate the disadvantages of having large test fixture mass between the force gages and the test item. Finally, new vibration test control systems need to incorporate the capability of specifying separate references for each control channel; a feature currently offered by only one major manufacturer.

ACKNOWLEDGEMENT

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

1. T. Scharton and D. Kern, "Using the VAPEPS Program to Support the TOPEX Spacecraft Design Effort," Shock and Vibration Bulletin, No. 59, October 1988.

2. A. Piersol, P. White, E. Wilby, J. Wilby, and P. Hipol, "Vibration Procedures for Orbiter Sidewall-Mounted Payloads," Astron Report on USAF Contract F04701-87-C-0010, Phase II Report, February 1989.

3. D. Smallwood, "An Analytical Study of a Vibration Test Method Using Extremal Control of Acceleration and Force," Proceedings of the IES, 35th Annual Technical Meeting, Anaheim, CA, May 1989.

4. T. Scharton, D. Boatman, and D. Kern, "Dual Control Vibration Testing," Shock and Vibration Bulletin, No. 60, November 1989.

5. T. Scharton, "Analysis of Dual Control Vibration Testing," Proceedings of the IES, 36th Annual Technical Meeting, New Orleans, LA, April 1990.

6. D. Smallwood, "Development of the Force Envelope for an Acceleration/Force Extremal Controlled Vibration Test," Shock and Vibration Bulletin, No. 61, October 1990.

7. T. Scharton, "Force Specifications for Extremal Dual Controlled Vibration Tests," Shock and Vibration Bulletin, No. 61, October 1990.

8. E. Skudrzyk, Simple and Complex Vibration Systems, Pennsylvania State University, University Park, PA, pp. 32, 1968. (Foster's Theorem brought to author's attention by D. Smallwood)

9. T. Scharton, "Dual Control Vibration Tests of Flight Hardware," Proceedings of the IES, 37th Annual Technical Meeting, San Diego, CA, May 1991.

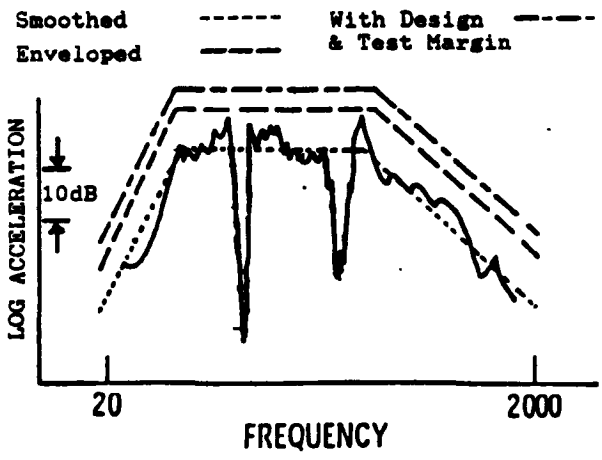


Fig. 1 Smoothing, Enveloping and Adding Margins to Flight Data

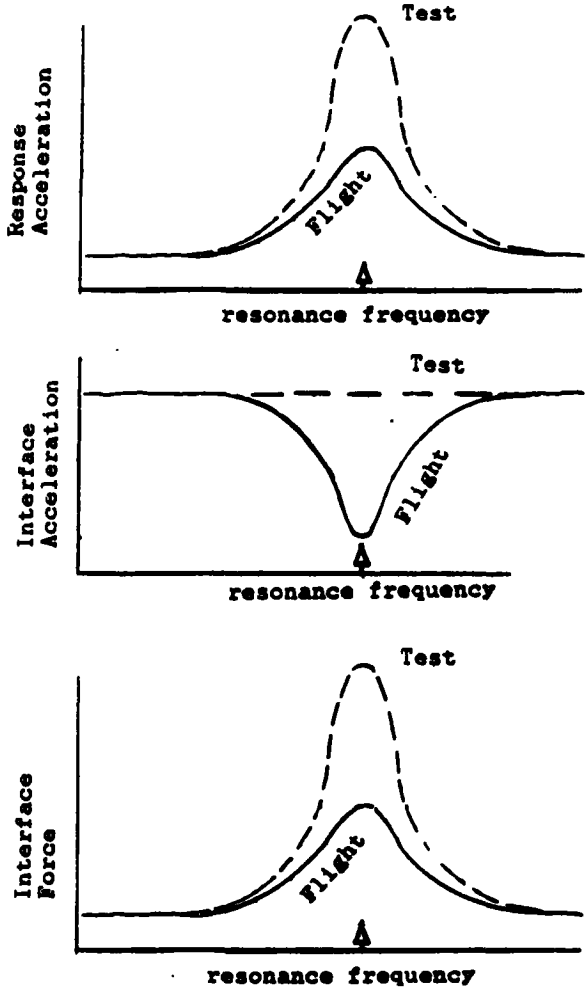


Fig. 2 Zoom Analysis of Force and Accelerations at Resonance of Test Item

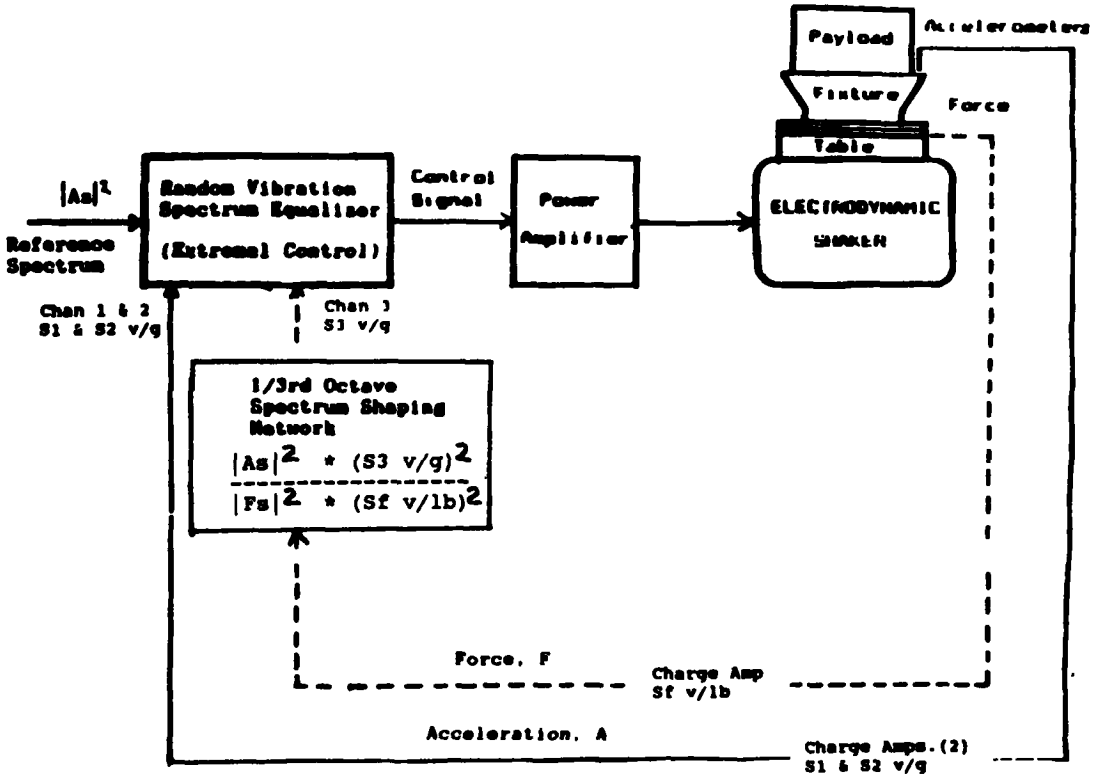


Figure 3. Diagram for Extremal Control of Acceleration and Force

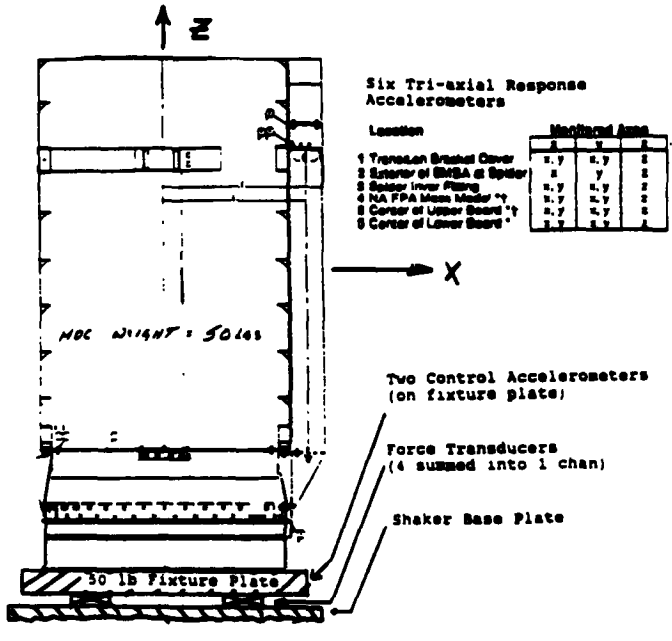


Figure 4. Mars Observer Camera Test Set-up for Vertical Axis Vibration Test

ORIGINAL PAGE IS
OF POOR QUALITY

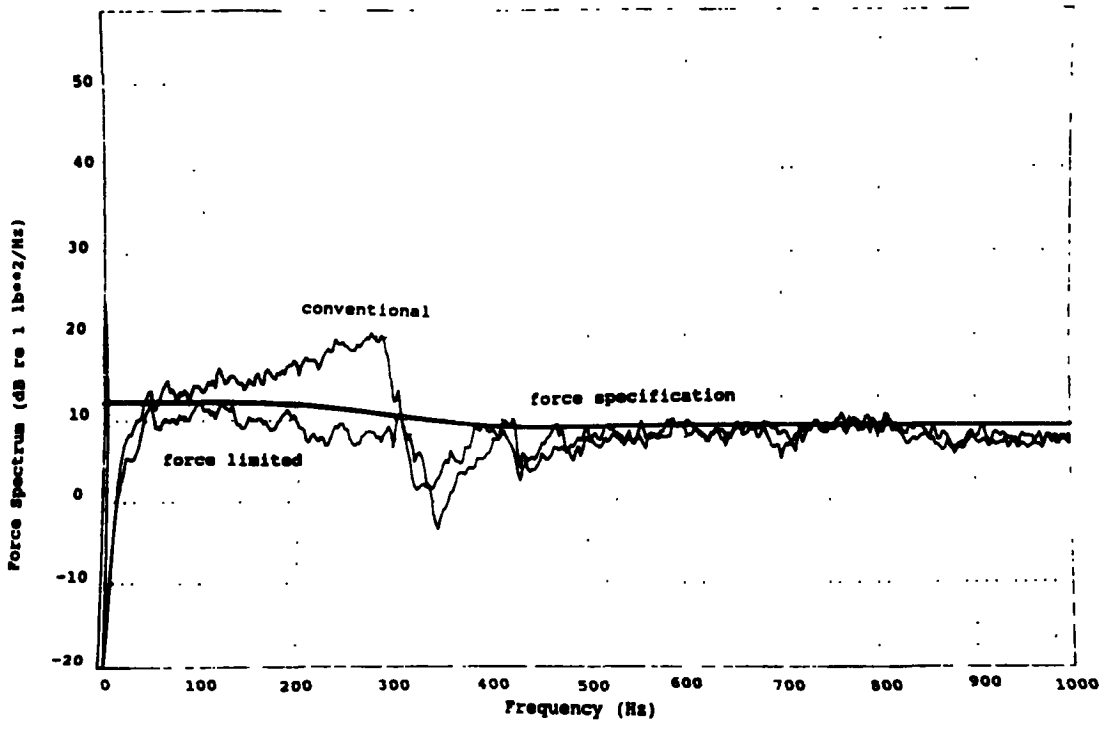


Figure 5. Force in Conventional & Force Limit Vert. -18 dB Random Tests

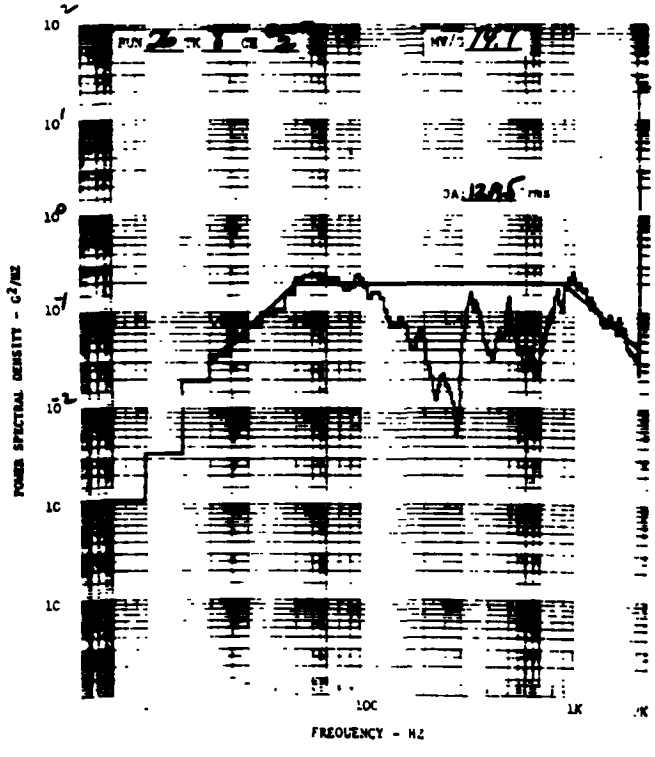


Figure 6. Input Acceleration in Force Limit Vert. Full Level Random Test

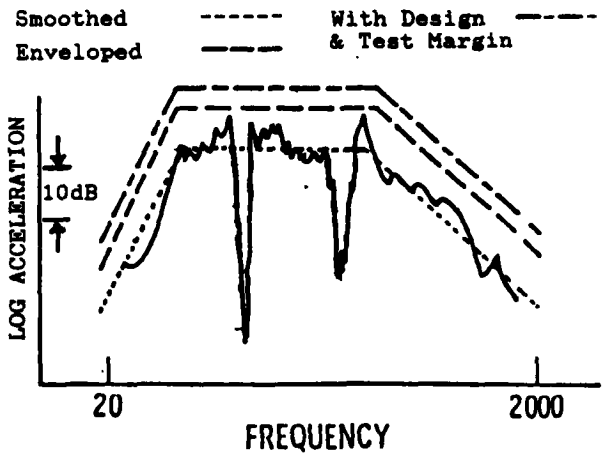


Fig. 1 Smoothing, Enveloping and Adding Margins to Flight Data

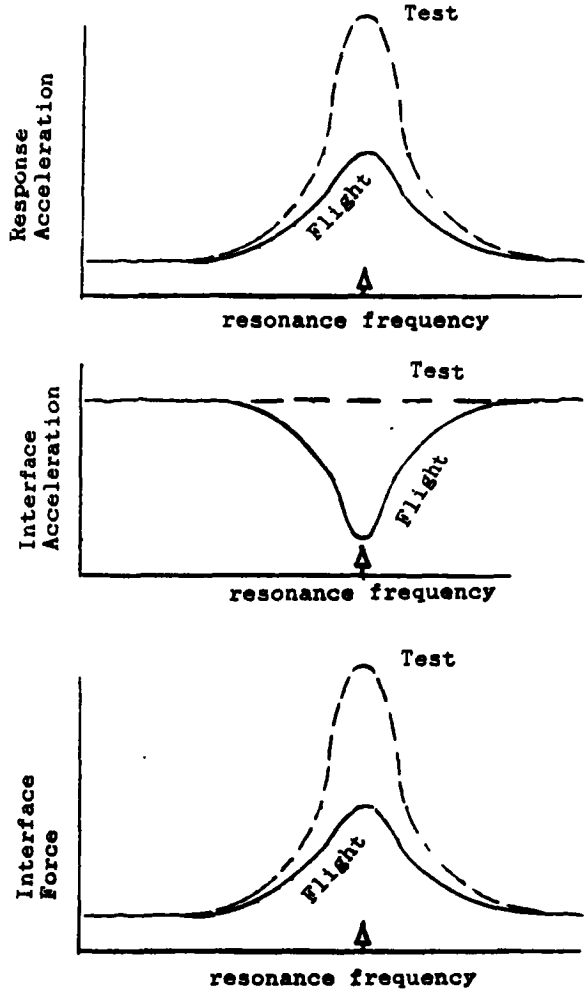


Fig. 2 Zoom Analysis of Force and Accelerations at Resonance of Test Item

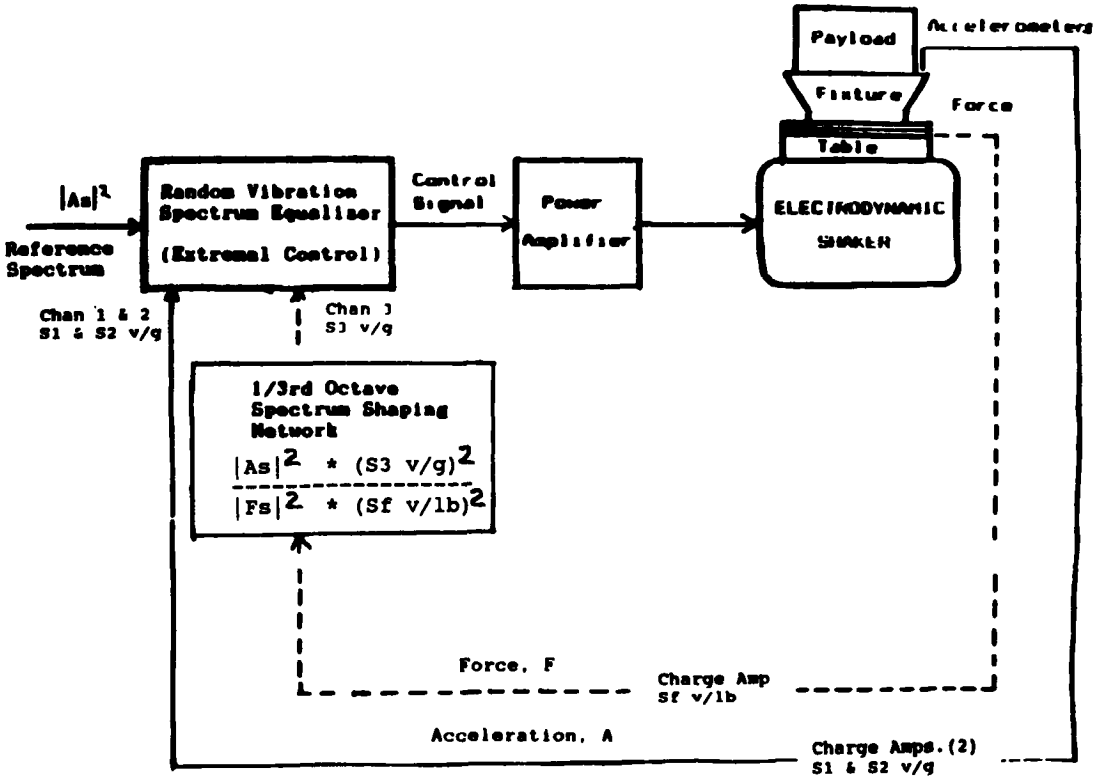


Figure 3. Diagram for Extremal Control of Acceleration and Force

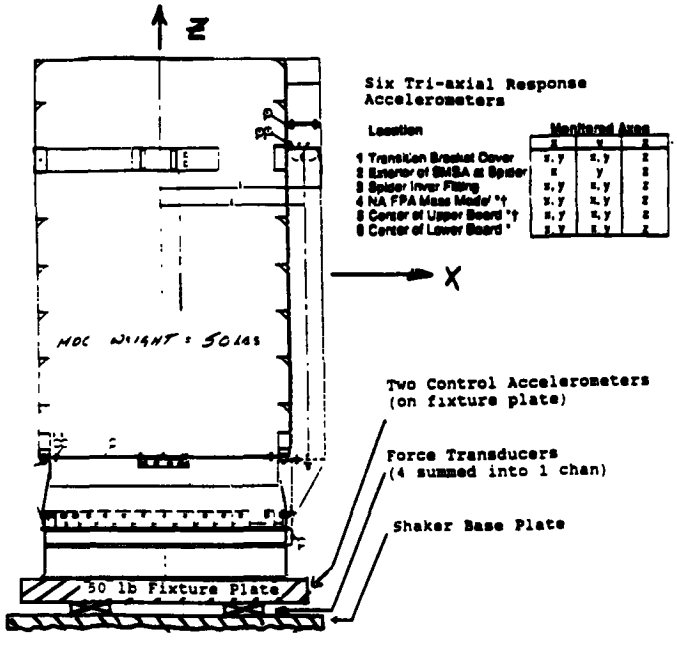


Figure 4. Mars Observer Camera Test Set-up
for Vertical Axis Vibration Test

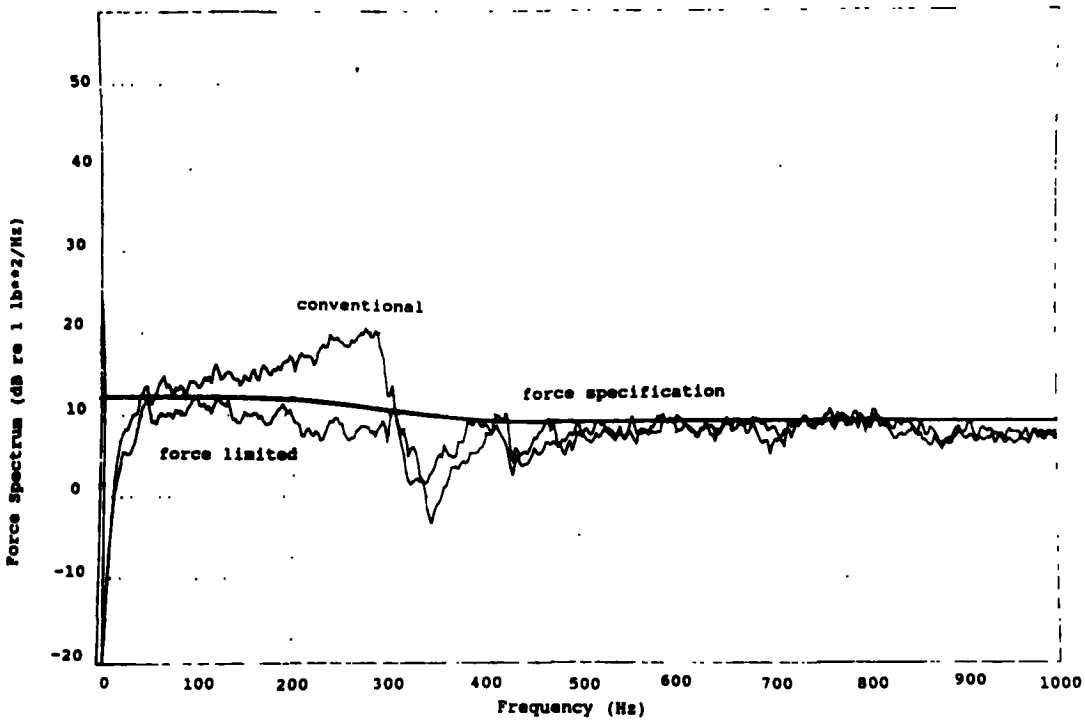


Figure 5. Force in Conventional & Force Limit Vert. -18 dB Random Tests

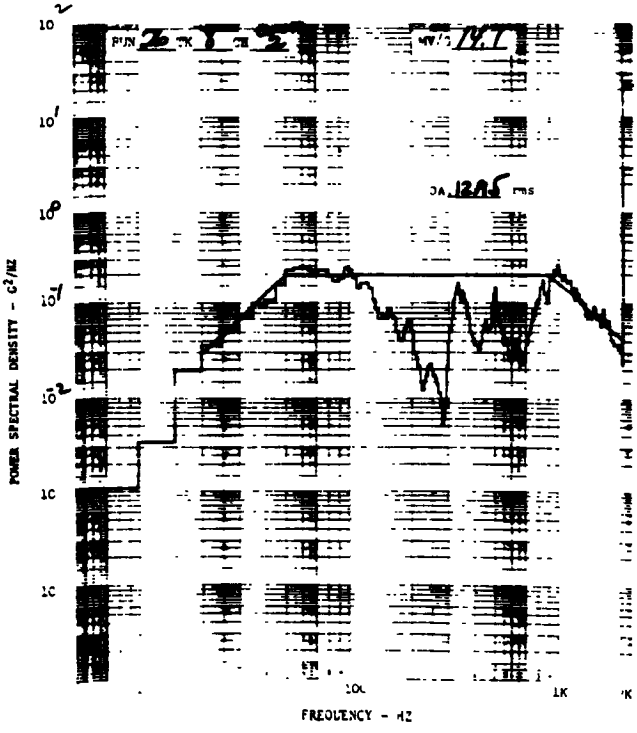


Figure 6. Input Acceleration in Force Limit Vert. Full Level Random Test

ADVANCED MANUFACTURING

(Session D1/Room A1)

Thursday December 5, 1991

- **Development of a Rotary Joint Fluid Coupling for Space Station Freedom**
- **Spline Screw Comprehensive Fastening Strategy**
- **Commercial Application of an Innovative Nut Design**
- **Inflatable Traversing Probe Seal**

PRECEDING PAGE BLANK NOT FILMED

N92-22689

DEVELOPMENT OF A ROTARY JOINT FLUID COUPLING FOR SPACE STATION FREEDOM

JOHN A. COSTULIS
NASA LANGLEY RESEARCH CENTER
HAMPTON, VA. 23665-5225

ABSTRACT

This paper describes the design and development of a rotary joint fluid coupling for Space Station Freedom. The function of this fluid coupling is to transfer a heat rejection fluid between stationary and rotating interfaces within the Space Station thermal bus system. The design of this coupling incorporates a modular type design to allow maximum flexibility, two types of seals (mechanical face seals and shaft lip seals) for redundancy, materials with excellent ammonia compatibility, and coatings that enhance seal wear resistance and increased protection against corrosion. This design has been thoroughly tested and has met all design requirements. Potential applications of this hardware include uses in gun turrets, coal slurries, farming equipment, and any other applications that require the low leakage transfer of fluids between stationary and rotating interfaces.

INTRODUCTION

The Space Station Freedom, as currently envisioned, will be gravity gradient stabilized and incorporate rotating solar arrays and thermal radiator panels. The rotation of the solar arrays and radiators, to track the sun and deep space respectively, will allow for a more efficient thermal and power system with a reduced overall size, weight, and cost.

This rotation requirement calls for the development of structural, fluid, and electrical mechanisms to provide for the transfer of loads, thermal fluids, and power across the rotating interfaces. As part of the Space Station Freedom advanced development effort, a R&D effort was begun at LaRC to generate a design of a rotary joint fluid coupling to transfer thermal fluids across a rotating interface within the Space Station Freedom thermal bus system.

The objectives of this advanced development program were to demonstrate the feasibility of rotary joints and to evaluate the design concepts proposed for 360° continuous rotation. These objectives included the identification of manufacturing and assembly methods, identification of candidate materials with ammonia compatibility, selection of seals, determination of leakrates, determination of drive power requirements, and lifetime assessment. From these objectives in conjunction with the performance requirements and the expected mission of the Space Station Freedom a set of design requirements were generated. These design requirements are as follows:

- 360° rotation capability
- 20 years of continuous rotation with little or no maintenance (100k revs.)
- .01 RPM rotation rate
- Anhydrous Ammonia compatible
- Flowrates based on 300 KW thermal load
- Leakrate less than .018 scc/sec (454 g/yr)
- 3 flow circuits (gas and liquid passages = 1 flow circuit)
- Less than 1.0 psi pressure drop per flow circuit
- Low drive torque

Three single circuit conceptual designs for rotary fluid couplings were investigated. An engineering model for each concept was designed and fabricated. Testing of each unit was then conducted in an ammonia flow testing facility which simulated the Space Station Freedom thermal bus. Based upon the results of the testing, an engineering prototype unit or Life Test Article (LTA), (see figure 1), was designed, fabricated, and successfully tested.

PRECEDING PAGE BLANK NOT FILMED

LTA ROTARY FLUID COUPLING DESIGN

A modular, systems type approach was utilized in the LTA design to seal the flowing ammonia. Mechanical face seals served as primary seals to isolate the flowing ammonia from the secondary rotary shaft lip seals (see figure 2). To accommodate the modular approach to sealing the ammonia, the LTA exterior housing was designed as segmented units. Segmenting the housing of the LTA provided the capability for both portions of the mechanical face seals and rotary shaft lip seals to be installed into the housing segments before assembly onto the shaft. Flexibility to increase the number of circuits was obtained by just adding additional segments. The housing segments were machined out of 6061-T6 aluminum alloy, and were sulfuric acid anodized which provided protection from the corrosive environment. Vapor and liquid flow annular passages were machined into the exterior housing segments. These passages were sized to provided for a larger hydraulic diameter which in turn decreased the pressure drop of the LTA. In addition to these passages, leak ports were machined into the housing segments in order to measure the leakrates of the primary and secondary seals. The outer two leak ports served as both leak ports and as scavenger ports to scavenge off any potential exterior leakage. The nonsegmented shaft which the exterior housing assemblies mount to was also machined out of 6061-T6 aluminum alloy. Vertical and horizontal flow passages were machined into the shaft. The shaft was coated with a General Magnaplate HCR coating, an anodized and teflon impregnated coating, which provided a hard, smooth, and low coefficient of friction surface for the seals to wear against. The outer diameter of the shaft was 3.0 inches. This provided enough space to accommodate all vapor and liquid flow passages, and allow for the use of standard, commercially available seal types and sizes.

The mechanical face seals incorporated 7 major elements (see figure 3). The metallic elements (spring, torque nut, and shell) were made of a 300 series stainless steel. The rotor (seal ring) was made of grade P8290 carbon graphite. The stator (housing seal) were made of grade PS9242 reaction bonded silicon carbide. The rotor and stator were lapped to a flatness of 1-3 Helium light bands. The diaphragm and o-ring seal on the stator were both fabricated of ethylene propylene.

Rotary shaft seals were used as secondary and backup seals in the LTA design. Ultrahigh Molecular Weight Polyethylene and 316 stainless steel were selected for the jacket and spring material of the rotary shaft seals (see figure 4). These seals provided redundant sealing capability for any leakage through the mechanical face seals. These seals also prevented the elastomeric components of the mechanical face seals from being exposed to the hard vacuum of a space environment.

ROTARY FLUID COUPLING TESTING

An ammonia flow test facility was designed and built to provide the range of flow and rotation rates required to simulate the Space Station Freedom thermal bus system. This system (see figure 5), was designed primarily to operate with liquid anhydrous ammonia. Flow rates from 0 to 2.65 gpm could be selected. A drive system provided rotation rates through the drive shaft from .01 to 1.00 rpm, with an output torque of 492 ft-lbs available over the entire range. The operating temperature of the system could be controlled from -35°F to 90°F using a separate temperature controlled liquid bath and heat exchanger.

The LTA was installed and tested in the ammonia flow facility. The pressure within the test facility flow circuit was maintained between 114 psig and 126 psig. This was controlled by maintaining the system temperature between 70 - 75°F. The rotation rate of the LTA was set at 1.0 rpm for the majority of the testing. This rate was somewhat arbitrary; however, the objective was to accelerate the test without modifying the performance of the seals. To obtain data at actual Space Station Freedom rotation rates, the drive speed was occasionally lowered to .01 rpm.

The LTA life testing performance parameters that were measured from this testing are as follows:

- Breakout torque
- Running torque at .01 & 1.0 RPM
- Flow circuit pressure drops
- Primary seal leakage at .01 & 1.0 RPM
- Secondary seal leakage at .01 & 1.0 RPM

- Exterior coupling leakage at .01 & 1.0 RPM
- Number of revolutions

The breakout and running torques were measured by strain gages mounted on the input drive shaft to the LTA. The drive torque produced by the LTA at .01 and 1.0 rpm was recorded on a strip chart throughout the life test.

The pressure drop through the flow circuit of the LTA was measured by a differential pressure transducer mounted between the inlet and outlet flow ports of the coupling. The pressure drop through the coupling was monitored and recorded continuously in order to determine the effect of shaft flow port position vs housing flow port position.

The primary and secondary seal leakages were measured by detecting the pH change in a controlled volume of a standard liquid after the liquid has combined with the ammonia that has leaked pasted the seals. Dry nitrogen gas was flowed through the primary and secondary seal leakports (see figure 6) in order to purge the ammonia into a tank containing the reference solution. The ammonia was then absorbed into the solution and the pH change is noted for that given time period. This solution was then titrated back to the reference pH value with a known amount of acid. Knowing the amount and type of acid required to titrate the solution back to the reference pH value, and the duration of the data period the seal leakage was calculated.

Exterior leakages from the end caps (Drive side and Loop side) of the LTA (see figure 6) were also derived in the same manner as the seal leakages. These leakages represent the leakage that would escape from the coupling or could be contained in some form of a scavenger system for on-orbit operations..

ANALYSIS OF TEST RESULTS

The LTA initial breakout torque was approximately 44 ft-lbs. After the initial breakout, the torque dropped to a level of 25 ft-lbs. As the seals proceeded to seat, the drive torque steadily increased to a maximum of 52 ft-lbs at 10,000 revs. Once the seals were seated, the torque began to steadily drop to a running torque level of 36 ft-lbs. This running torque level was observed for the remainder of the test (200,000 revs.) . No significant change in drive torque was observed when the rotation rate was lowered to .01 rpm. (see figure 7)

The LTA end cap leakage was the amount of ammonia that leaked past the end seals on either side of the coupling. The drive end cap leakrates ranged from $8x10^{-3}$ to $1x10^{-3}$ scc/sec at 1.0 rpm and $5x10^{-3}$ to $1x10^{-3}$ scc/sec at .01 rpm. The LTA loop end cap leakrates were in the same range as the drive end side.

The secondary seal port leakage was the amount of leakage that leaked past the secondary rotary shaft lip seals. The secondary seal leakage ranged from 4×10^{-2} to 1×10^{-3} scc/sec. The primary seal port leakage was the amount of fluid that leaked past the primary mechanical face seals. The primary seal leakage ranged from 4×10^{-2} to 1×10^{-2} scc/sec.

Pressure drop through the coupling was measured and recorded on the facility data acquisition systems. The data showed pressure drops from .35 to .50 psid. This variation in pressure drop depended on the orientation of the shaft flow port relative to the flow port in the exterior housing. The pressure drop was cyclical with each revolution of the shaft.

CONCLUSION

The LTA met all performance requirements for a 20 year equivalent life test (100,000 revolutions). This coupling was actually tested to twice the expected mission life (40 yrs. or 200,000 revs.). Primary seal leakrates for the coupling were in the 10^{-2} scc/sec range and secondary seal leakrates were in the 10^{-3} scc/sec range. Running

torque for the coupling was 35 - 37 ft-lbs, and breakout torque was approximately 45 ft-lbs.

With slight modifications to the baseline design, the exterior (scavenger port) leakage can be reduced further to the 10^{-4} scc/sec (.01 lbm/yr) range. This modification would entail the addition of 2 rotary shaft seals located on either end of the coupling. These additional seals will increase the drive torque slightly, but will drastically reduce the size of the ammonia scavenger system.

The versatility of this rotary fluid coupling design will allow for its use in a variety of applications. This design can easily be modified to accommodate different shaft and housing materials, a variety of flow circuit configurations, pressures, and temperature ranges.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

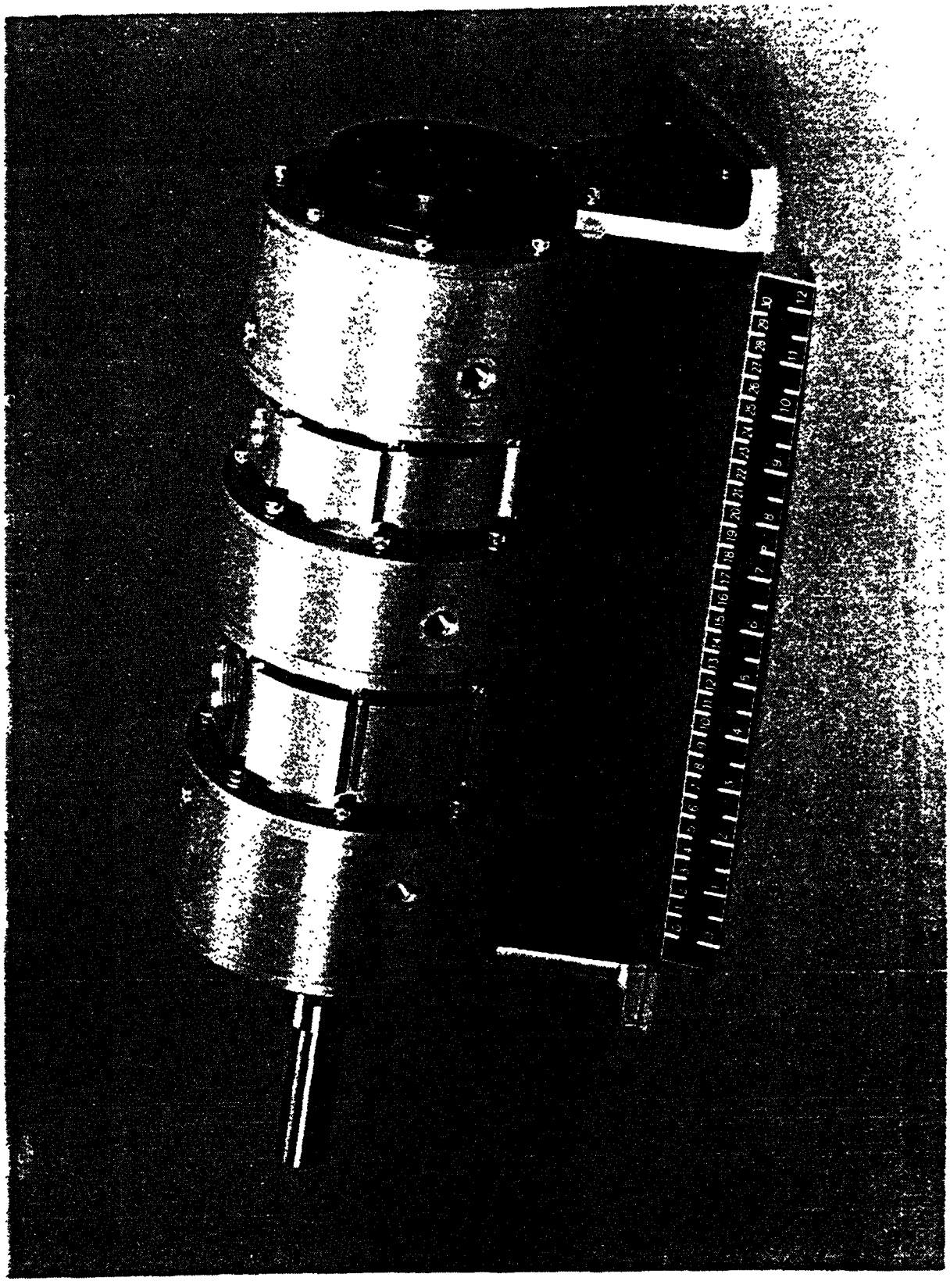


FIGURE 1: LIFE TEST ARTICLE

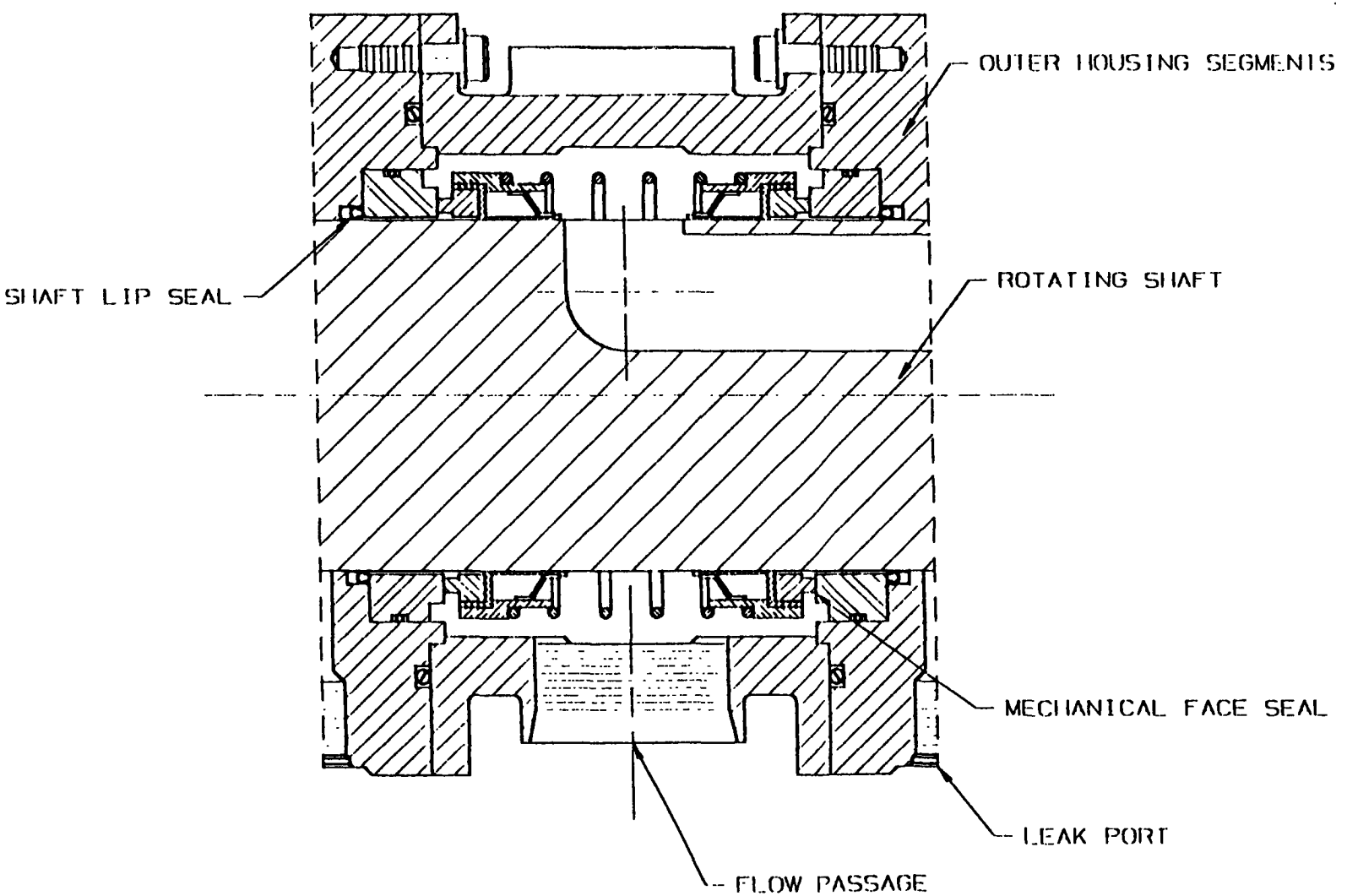
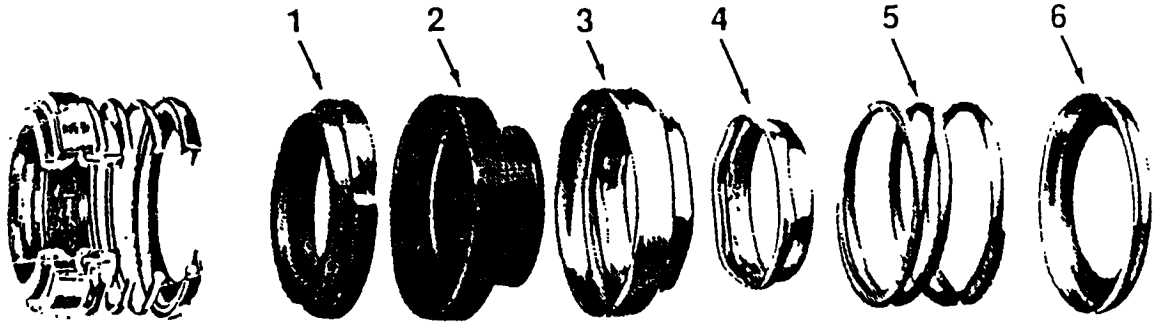


FIGURE 2: TYPICAL SEAL ASSEMBLY



1. Seal ring
2. Diaphragm
3. Shell
4. Torque nut
5. Spring
6. Shaft seal
7. Housing seal

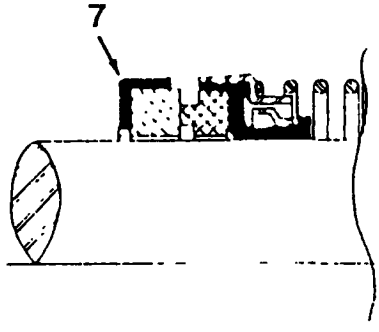


FIGURE 3: MECHANICAL FACE SEAL ASSEMBLY

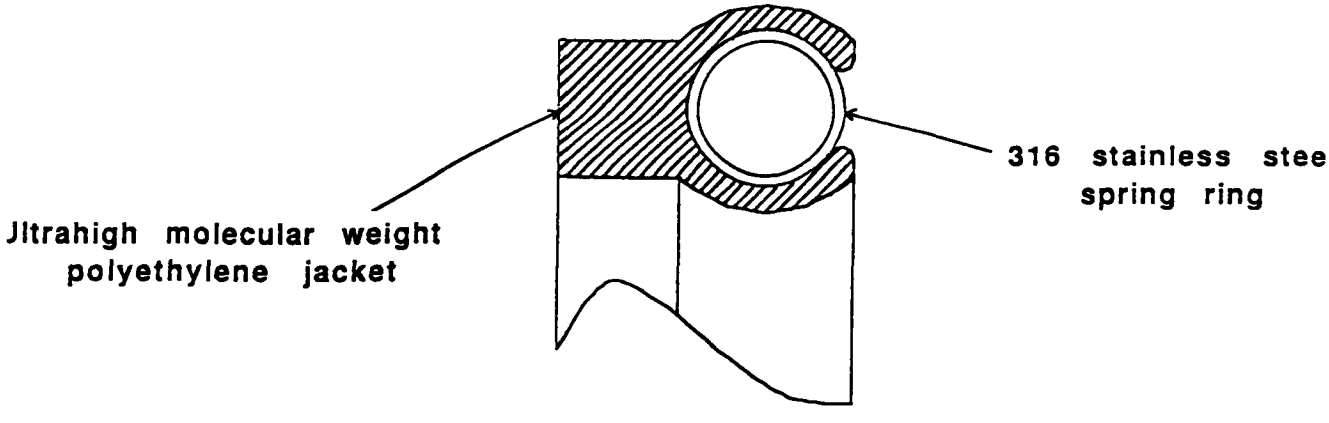


FIGURE 4: ROTARY SHAFT LIP SEAL

FIGURE 5: ROTARY JOINT TEST FACILITY

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

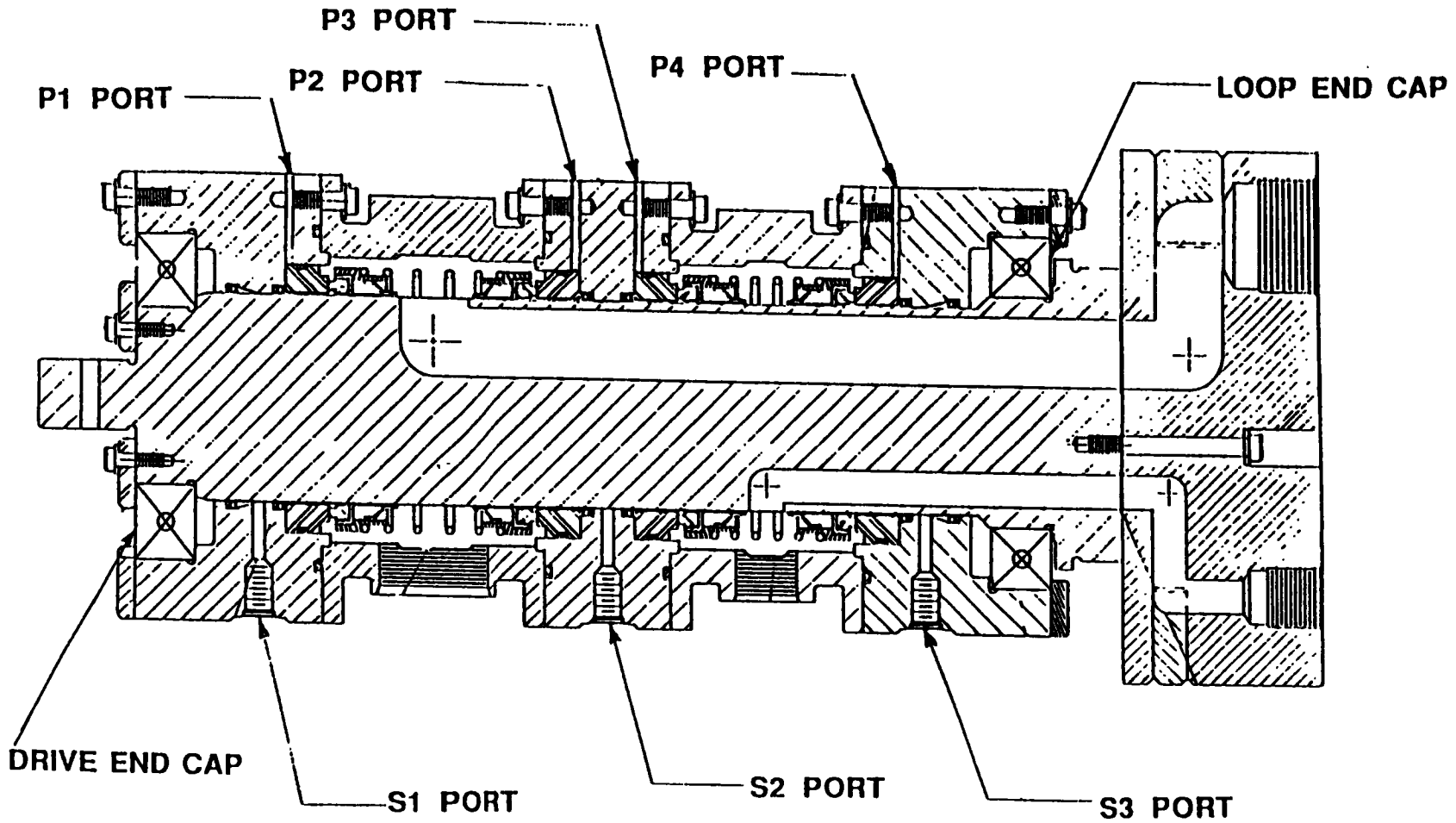


FIGURE 6: LTA LEAK PORTS

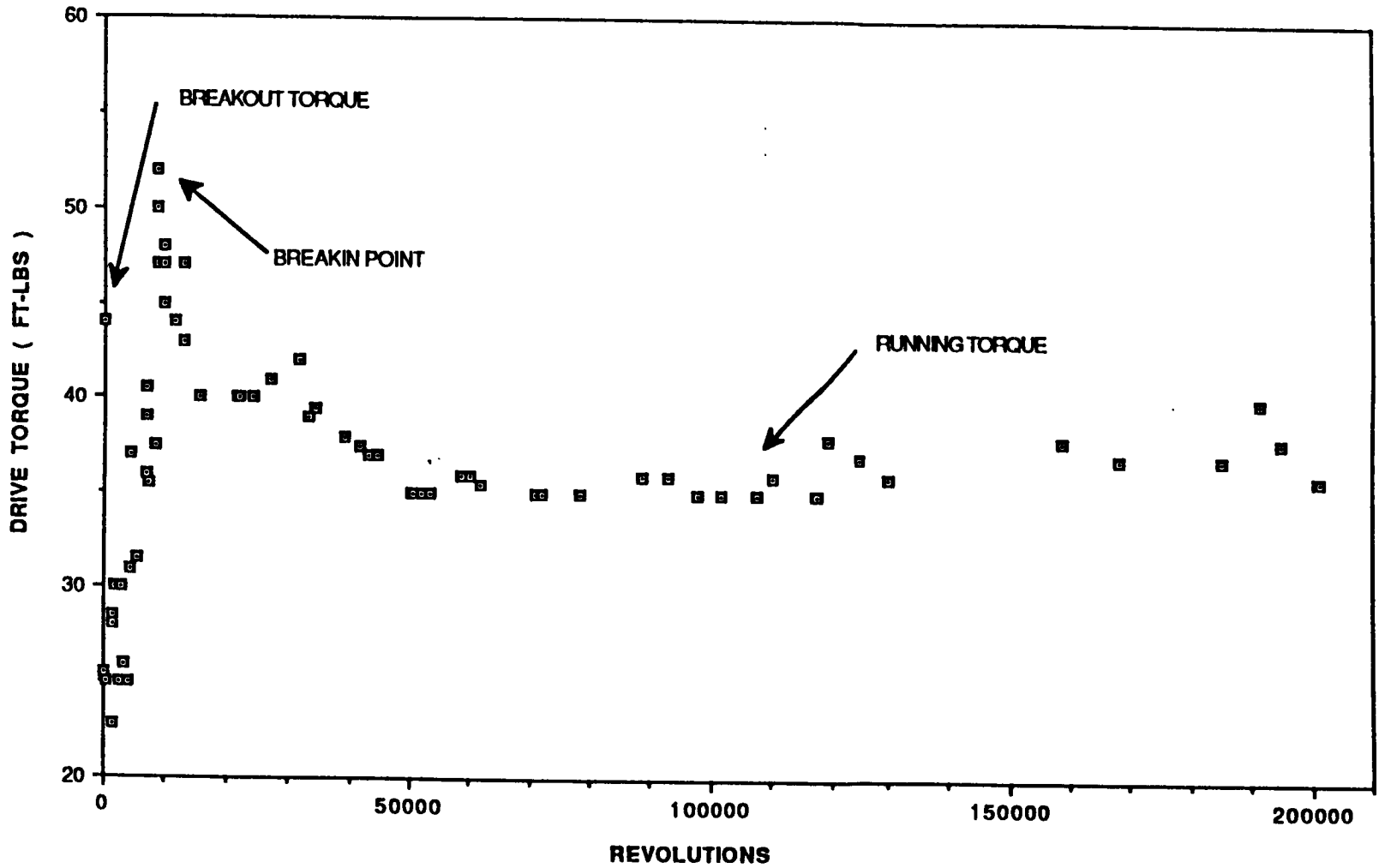


FIGURE 7: LTA DRIVE TORQUE

N92-22690

SPLINE-LOCKING SCREW FASTENING STRATEGY (SLSFS)

John M. Vranish
NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

A fastener has been developed by NASA/GSFC for efficiently performing assembly, maintenance, and equipment replacement functions in space using either robotic or astronaut means. This fastener, the "Spline-Locking Screw" would also have significant commercial value in advanced manufacturing. Commercial (or Department of Defense) products could be manufactured in such a way that their prime subassemblies would be assembled using Spline-Locking Screw fasteners. This would permit machines and robots to disconnect and replace these modules/parts with ease, greatly reducing life cycle costs of the products and greatly enhancing the quality, timeliness, and consistency of repairs, upgrades, and remanufacturing.

The operation of the basic Spline-Locking Screw fastener will be detailed, including hardware and test results. Its extension into a comprehensive fastening strategy for NASA use in space will also be outlined. Following this, the discussion will turn towards potential commercial and government applications and the potential market significance of same.

INTRODUCTION

In space operations, fastening problems are unusually important. Common machine screws cannot be applied with the same ease as on earth; they cross-thread easily because the astronauts must wear gloves and space suits when performing maintenance. Robots have cross-threading problems and more. At the same time, the violent vibrational and loads environment generated by launching payloads into orbit mandates a requirement for strong, simple, light-weight, reliable fasteners that must be met. The Spline-Locking Screw, described in this paper addresses this problem directly. But; in developing what amounts to a screw that cannot be cross-threaded, it soon became apparent that more was involved. One could add alignment and torque reaction pins and accomplish precision and reliable electrical connections in addition to the fastening by making slight modifications to the basic Spline-Locking Screw. A few modifications more and a Standard Robot End Effector and related Astronaut Hand Power Tool emerged. This Standard Robot End Effector (or Astronaut Hand Power Tool) could acquire other power tools or store them in holsters as required. It thus became clear that, using this approach, any number of complex (or simple) tasks could be accomplished in sequence by progressively modifying a basic Spline-Locking Screw system. The purpose of this paper is to provide an introduction to the Spline-Locking Screw concept and its derivative devices and to give an indication of its potential for a comprehensive fastening strategy ; both for space and commercial applications.

A "proof-of-principle" prototype of a foot to fasten the "Flight Telerobotic Servicer" (FTS) robot to Space Station structure based on the Spline-Locking Screw concept has already been built and tested (Fig.1). The results of this effort have immediately shown success and the device has been adopted by the prime contractor and NASA as the foot that will be employed on the FTS robot when it is developed for flight. This particular device, known in NASA as the Workpiece Attachment Mechanism/ Workpiece Attachment Fixture (WAM/WAF) (see Fig. 1) does electrical connections in addition to basic fastening. It will also, of course, permit a robot to walk on a space structure.

THE SPLINE-LOCKING SCREW

In this section the evolution of the old NASA fastener used when repairs in space were anticipated will be outlined along with the problems this evolution caused. This will set the stage for the development of the basic Spline-Locking Screw concept.

The evolution of this NASA fastener [1] is shown in Fig. 2. [Many of these lessons were learned as part of the Solar Max satellite repair mission of 1982]. Screws with low pitch machine threads had been used

successfully many times in space launch operations. But; these would cross-thread when the Astronauts tried to fasten them while wearing gloves and space suits. A guide was added to the screw and this helped; but not enough. Next, the low pitch machine thread was changed to a more high pitch acme thread. This worked; but now the screw would back out during vibrational tests simulating launch. To overcome this, a taper interface at the top of the screw was added. At this point everything worked. However, the additions added a great deal of friction to the system and this meant that large torques on the order of more than 100 ft-lbf. would be required to free up the bolts used in the Solar Max mission. This, in turn, meant that very large and clumsy hand tools had to be used and the Astronauts (or robots) could be subjected to dangerously large torques and forces. For a robot, the situation was nearly impossible. Using the large tools would severely limit robot dexterity. On the other hand, the existing robot End Effectors could only produce 20 ft-lbf. torque[2].

The Spline-Locking Screw was developed by returning to the common machine screw and taking a fresh approach[3,4]. It was decide to cut the bolt in two and leave the bottom half threaded (Fig. 3a.). A new interface was created in the shaft of the bolt. Thus, we have the bolt head (or Driver) and the Object which is normally pinched by the screw system on one side of the interface and the Bolt and the threaded Fixture ,it screws into, on the other. The problem, then,was to create an interface that would complete the system. A Spline -Locking type interface (Fig. 3b.) was chosen because it was simple, direct and effective. There is an increase in size; but it is minimal. The operational concept of the system is shown in Fig. 4. The Driver is inserted into the Bolt such that the Male Splines of the Driver fit into matching Female Splines of the Bolt. The Driver is then turned clockwise to tighten the screw system. During this process, the male Driver Spline engages in the female Bolt Spline, it cannot be pulled out and the Bolt and Driver turn together as a single complete machine screw. To unfasten the system, the Driver is turned counter-clockwise, the screw loosens and the Driver Splines relocate in the Bolt (disengage) such that the Driver can easily be pulled out of the Bolt. The Splines are very coarse so they can be designed to seat and fit together such that cross-threading is virtually impossible. And the Bolt is never unthreaded, thus, we have, in effect, a machine screw that cannot cross-thread. At the same time, it is now possible to use low pitch machine threads so large preload forces can be generated from minimum input torques (on the order of 8 ft-lbf. to generate 1,000 lbf. preload). And, these Bolts would not shake loose during launch because of their low pitch. Also, robot End Effectors and Astronaut Hand Power Tools could now be made more modest in size and power, enhancing safety for the Astronaut and safety and dexterity for the robot.

As a practical matter, the Spline -Locking design requires careful and detailed treatment. This is to make certain that the Splines engage and disengage properly and that the payload remains attached either to the Robot End Effector (or Astronaut Power Hand Tool) or to the Fixture at all times to prevent it drifting off into space. We will begin examining these details by following the Tightening sequence of steps shown in Fig. 5a. When the Driver encounters the Bolt, it normally initially comes to rest on the top of the Bolt. As the Driver is rotated clockwise, it rotates with respect to the Bolt until the Driver and Bolt Splines line up; the Driver Bias Spring pushes the Driver into the Bolt (Driver Spline Insertion, Fig. 5a) and the Driver is seated in the Bolt. This raises a question as to how we can be certain the Bolt will not turn with the Driver and prevent the relative motion of the male and female Splines essential to insertion and seating. This is accomplished by adding a Preload Spring to the Bolt (Fig. 3b) to ensure that the Bolt will not turn until the Driver Spline drops into the Bolt Spline, is seated and the "Splines Engaged" (Fig. 4b and Fig. 5a). At this point the Driver torque will simply overcome the friction from the Preload Spring and the Driver and Bolt will turn together. With the Driver and Bolt turning clockwise together, the Bolt will translate downwards and apply a large Locking Force between Driver and Bolt Splines. A subtle distinction is involved between the terms "Spline Engagement" and "Spline Locking" which can be seen by comparing Figs. 4 and 5 and noting that an Underhook Region has been added to the female Bolt Splines. The reason for the Underhook region will be made clear below.

The details of loosening will now be discussed (Fig. 5b). The sequence starts with the Splines locked together and the Object preloaded to the Fixture (Fig. 3). As the Driver is turned counter-clockwise, the Bolt Spline is held in the Driver Spline by the Underhook (Fig. 5b). Thus, the Bolt and Driver must turn together to break the Bolt Thread loose and to release the preload force. As the counter-clockwise rotation continues, the Driver remains in its downward position because of its Preload Spring, but the Bolt translates upwards (Fig. 5b). This causes the Splines to unlock and reposition for Spline Disengagement and Removal. As the Bolt translates upwards, it is capable of generating a large force to "push" the Object away from the Fixture. This condition is termed "Push-Off" and prevents cold welding or jamming of the

Object to the Fixture. At this point, the Driver and Bolt turn together until the Bolt threads hit a stop. The Driver Splines and the Bolt Splines remain aligned throughout for easy removal of the object from the Fixture. Some of the reasons for the Underhook Region now begin to emerge. This addition makes certain that, during the unfastening process, the Bolt is located in the Fixture in the same position each time before the Driver and Object can be removed. This also makes certain that the Bolt and Driver are properly positioned to begin the Insertion and Fastening process shown in Fig. 5a.

ELECTRICAL CONNECTOR CAPABILITY

In this section, it will be shown that an electrical connection capability can be added to the fastening capability by adding minor modifications to the basic Spline- Locking Screw[4]. Further, it will be shown that this leads to an entirely new set of devices to include End Effectors with Tool Storage and Tool Autochanger capabilities, the WAM/WAF and an Astronaut Power Hand Tool.

An electrical connector capability can be incorporated in the Spline- Locking Screw concept as shown in Fig. 6. The Spline-Locking Screw Bolt would be threaded on a Nut rather than into the Fixture. The Bolt is coupled to the Fixture by means of a Preload Spring F_S (nominally 100lbf) forcing the Bolt down towards the Fixture. Contact between the Bottom Stop on the Nut and the Fixture prevents the downward translation of the Bolt. Thus, the Bolt and Nut are preloaded against the Fixture with a force equal and opposite to that of the Preload Spring (labelled $F_{RS\,1}$ in Fig. 6). Also, the interface between the Nut and Fixture are splined so that the Nut cannot rotate; but can translate between the Fixture Top and Bottom Stops. Electrical Pins can be added to the Nut, and Pin Receptors added to the Fixture. In this section, both the tightening and loosening sequences will be examined.

As the Driver turns it first seats in the Bolt and follows all the steps associated with the tightening process described above through Spline Locking. At this point, the $F_{RS\,1}$ force transfers from the Fixture/Nut interface to locking the Driver and Bolt Splines together (shown as F_D in Fig. 6). The F_D force, in turn, is reacted by the equal and opposite force (labelled $F_{RS\,2}$ in Fig. 6 forcing the Fixture against the Object. All the above forces $(F_S, F_{RS\,1}, F_D, and\ F_{RS\,2})$ are equal to each other. They are given different subscripts because they exist at different times and at different locations in the fastening sequence. As the Driver continues to turn clockwise, the Bolt turns with it and, since the Bolt cannot translate downwards, being held in place by the Locked Splines of the Driver and Bolt, and the Nut cannot rotate, the Nut translates upwards. Throughout this process, the force sustaining the Spline- Locking remains constant and equal to that of the Preload Spring. And, The Object and Fixture are forced against each other with the same force. We thus, have a new condition which will be termed ***Hard Dock***.With the proper Alignment Guides on the Fixture and Object (not shown in Fig.6), proper preconditions have been achieved for the electrical connection. As the Driver continued to turn clockwise, the Nut translates upwards until a precision electrical connection is made. Shortly afterwards, the Nut hits its Top Stop on the Fixture. With Nut translation stopped, the Bolt once again attempts to translate downwards. This forces the Object and Fixture together with preload forces. This condition is termed ***Preload***. Once again, it should be noted that Spline Locking can be done in either a ***Hard Dock*** or ***Preload*** condition, depending on the circumstances. Both conditions are useful in ensuring that the Object is properly secured to the Fixture throughout the Fastening process. It is also, perhaps, appropriate to note that the Object is secured (or Docked) to the Fixture as soon as the Splines are Engaged Fig. 5a), prior to Hard Dock. However, there is rattle between the Object and the Fixture during this condition so we will describe it with the term ***Soft Dock***.

Turning the Driver counter-clockwise reverses the steps described above and disengage the Object and the Fixture as well as the Driver and the Bolt. As previously described, under Hard Dock and Preload conditions the Driver Spline is seated in a groove in the Bolt Spline (Fig. 5). As described above, during counter-clockwise rotations, the Bolt Spline Underhook provides an interference obstacle preventing the two Spline Sets from slipping out of engagement. It should be noted, that during the disengagement of the Electrical Connectors, the Bolt Preload Spring provides the disengagement force and the force holding the Splines in Lock (and Hard Dock) is the difference between the Bolt Preload Spring force and that used in pulling the Electrical Connectors apart. And, this Hard Dock condition remains until the Nut bottoms and the Bolt

translates upwards, taking the system out of Hard Dock, unlocking the Splines, and into Soft Dock. Again, during this transition from Hard Dock to Soft Dock, the Bolt translates upwards and pushes the Fixture and the Object slightly apart (say 0.060 in.). This condition is termed Push- Off. The Bolt Spline is now free of the Under Hook and so the Driver Spline rotates with respect to the Bolt Spline until it hits a stop and both Driver and Bolt are stopped. This forces the two Splines to be lined up such that they can easily and reliably be pulled apart.

DERIVATIVE DEVICES

It would now be possible to make further minor modifications to the Spline- Locking Screw and produce a range of useful devices. One such device would be a combination Special Tool Interface and Autochanger [5]. This is a straightforward extension of the arrangement shown in Fig. 6. The Object of Fig. 6 can be fastened to a robot wrist and a motor splined to the Driver. This Coupling between Motor and Driver would include a Compliant Spring to permit the Driver to be pushed up out-of-the way. This arrangement of Object, Electrical Pin Receptacles, and Compliant Driver with Motor would constitute a Standard End Effector. Alignment Pins would be added to the Fixture (which would also serve as the Special Tool Interface) and Mating Receptacles would be incorporated in the Standard End Effector. Any Special Tool could be fastened to the Interface and the robot. Thus the robot could use the Standard End Effector to mate with and fasten to a common Special Tool Interface. And, since any Tool could be attached to the Special Tool Interface, and power and signal provided, the robot could acquire and use any of a wide variety of Tools. It should be noted that the mating procedure by which the robot acquires a Special Tool would use standard practice. That is, alignment is standard "peg-in-the-hole" using alignment pins and mating receptacles. (The WAM/WAF (Fig. 1)[4] is essentially a large version of this device which has the strength to withstand torques and forces on the robot leg.) The bottom of the Nut can be used to pinch tabs of a Tool Storage Holster and there by store Special Tools[5]. This system would be equally effective on earth or in Zero or micro "g". Thus the system would become its own Autochanger. Although not shown in this paper, a similar approach could be taken in permitting a robot to use "Spline-Locking Screw" techniques to release and fasten payload boxes known in NASA as Orbital Replacement Units (ORU)[6].

CALCULATIONS AND TEST RESULTS

A WAM/WAF prototype (Fig. 1) was constructed and tested as a first prototype. The prototype has been tested and demonstrated on a robot and found to dock, and to go through its proper fastening sequence of Soft Dock, Hard Dock, Electrical Connections (to include actuating dust covers on both WAM and WAF), and to provide Preload Forces sufficient to allow the robot to wave a large steel table around with impunity. Without question, it has great holding strength with minimal motor torque required. More detailed testing is being conducted in the Goddard Space Flight Center robotics lab. Calculations indicate that the WAM/WAF can produce excellent preload forces with modest actuation torques.

$$dW_{in} = dW_{out} + dW_{losses} \tag{1}$$

dW_{losses} = friction losses in bolt and in reaction thrust bearing

$$T(\theta)d\theta = F(\theta)\{\frac{d\theta}{2\pi L} + u_{s1}R_1 d\theta + u_{s2}R_2 d\theta\} \tag{2}$$

$$T(\theta) = F(\theta)\{\frac{1}{2\pi L} + u_{s1}R_1 + u_{s2}R_2\} \tag{3}$$

Where:

$dW_{in} = d$ if *ferential work done by the Driver on the Bolt*.

$dW_{out} = d$ if *ferential work done by the Bolt as it translates.*

dW_{losses} = *friction losses in Bolt and in reaction thrust bearing during* d if *ferential work.*

$T(\theta)$ = *input torque*.

$F(\theta)$ = *Bolt preload*.

u_{s1} = *Nut – Bolt friction coefficient* = 0.15.

u_{s2} = *coefficient of friction of rolling friction* = 0.008.

R_1 = *Bolt radius.*

R_2 = *radiusof thrust bearing.*

$d\theta$ = *d if ferential twist angle.*

L = *Bolt thread lead.*

Thus a 0.75 in UNF 16 bolt producing 1000 lbf preload requires
< 6.5 ft-lbf input torque which is a very modest value.
We will now examine the efficiency of the system.

$$E = \frac{dW_{out}}{dW_{in}} \tag{4}$$

$$E = \frac{F(\theta)}{2\pi L T(\theta)} \tag{5}$$

For the bolt, preload and coefficient of friction listed above, we get an efficiency of 12.8% which is more than satisfactory.

The WAM/WAF will not back drive and hence a brake is not required.

$$\tan(Y) = \frac{1}{2\pi L R_1} \tag{6}$$

Where y = Bolt lead angle.

When $\tan(Y) < u_{s1}$ *system will not back drive.*

$$u_{s1} 2\pi L R_1 = \text{safety factor} \tag{7}$$

For our WAM/WAF, we get a safety factor>5 so clearly a brake is not required.

COMPLEX ELECTROMECHANICAL SYSTEMS

A few more simple modifications can be added to the Locking Spline Screw to produce a multi-rotational output[7] which can be monitored by electronic signals throughout the process. This capability will, in turn, form the basis for using Locking Spline Screw techniques to operate complex electromechanical systems. Once again, some events must be done serially. During the process in which the robot acquires the Object, we have a sequence of: mating ; Soft Dock; Hard Dock; Electrical Connections; Multi-Rotational Output until a Stop is reached and the Object is released to the custody of the robot; then, finally, Preload. During the process in which the robot fastens an Object to some Fixture, we have a sequence of: Release of Preload ; Hard Dock; Multi-Rotational Output until a Stop is reached and the Object is fastened to a Fixture; Release of the Electrical Connectors; Soft Dock; Push Off of the robot Standard End Effector from the Object; and Separation of the robot Standard End Effector from the Object. It is apparent that with multiple rotations available from an output shaft which will turn until it reaches a stop, any number of different types of electromechanical systems can be driven by the shaft including items as complex as clocks. We would have, then, what amounts to a portable motor, controller , power supply, and system software and electronics and mechanical interfaces. All that most particular applications would require is an embedded mechanical system and sensors. This would vastly simplify many tasks. And, in those instances where an embedded motor is still required, the robot could supply power and controls.

SMALL OBJECTS

The examples above have shown that there is virtually no limit to the complexity of tasks that can be done by the Spline-Locking Screw system. This system could also handle extremely Small Objects (2 in.X2 in.X1.5 in.)(5 cm.X 5cm. X 3.8 cm)[6]. Handling Small Objects in space would be more formidable than is commonly realized. The main problem is the micro "g" environment that requires that every Object be fastened to something or it will float away (or worse-accelerate to missile-like speeds). This requires that control be fastidiously maintained during each step of the hand-off process despite the small size of the Object and the required simplicity of the fastening scheme. Adding to the difficulty is the requirement that the small object be grasped by the Spline-Locking Screw Standard End Effector". The problem can be

solved by piercing the Small Object with a Bolt, one end of which is a Driver interface and one end of which is the Bolt Spline (female) interface which mates with the Driver of the Standard End Effector. A small Rotating Socket with compliant spring would be embedded into the Fixture. Alignment Tabs would be placed in the Object, the Robot End Effector and the Fixture to permit the Small Object to be properly aligned in rotation at all times. The Bolt would be threaded into the Small Object so that as the Robot Driver turned, the Small Object Bolt would turn and translate up or down and, in the process, screw the Object off the End Effector and on to the Fixture or off the Fixture and on to the End Effector as required. The Small Object would be in proper control at all times.

FASTENING STRATEGY FOR SPACE OPERATIONS

The discussion above has shown the unusual capabilities of the Spline- Locking Screw approach. Further, it is clear that this fastener could form the basis for performing a host of operations ranging from attaching small, simple objects to acting as a transportable motor/control system/power supply and systems interface for complex electromechanical systems to permitting a robot to walk on a space structure or handle tools with the same appendages as the case may be. And, all of these devices could be actuated by the same rotary Driver (Identified as The Robot End Effector when used with robots or the Power Hand Tool when used by Astronauts). It would seem that a proper frame work has been laid for a comprehensive fastening strategy in which Spline-Locking Screw mechanical and electrical interfaces can be standardized into a few size ranges (like industrial machine screws) and any of a number of applications, techniques and innovative designs could be implemented consistent with those interfaces. Parts counts on space craft would be drastically reduced, modularity would be enhanced and maintenance and repair greatly facilitated. Robotics could now be employed more extensively in assembly and maintenance operations and Astronauts would also find things simplified and safer.

COMMERCIAL APPLICATIONS

Humans using their bare hands do not normally cross-thread bolts. Thus, from an industrial stand-point, Spline-Locking Screw would be another of many available fasteners serving a small, but important, niche' market , typically involving bolts of 0.5 in. dia. or larger in which the object they are fastening must be periodically removed and replaced using power tools. With these larger bolts, threads are course, the torque involved in installing them is large and, because of their size, Locking Splines can easily and cost effectively be employed. Because fine machine threads could be used with the Spline-Locking Screw, required torque would be reduced and with it the size of the power tools and objects would be held more securely against working their way free during vibration. Also, because the Spline-Locking Screw Bolt is pre-threaded, attachment is very quick; just a few turns. In the airline industry, for example, aircraft engines could be dropped quickly , overhauled and remounted. Similarly, the avionics would be installed using the Spline-Locking Screw with electrical connector. This, in turn, suggests that the computer industry could make extensive use of Spline-Locking Screws with electrical connectors. Earth moving and materials handling equipment could profitably use such fasteners extensively as could the automotive industry (wheel lug nuts and engine and transmission mountings come to mind immediately). Military applications are particularly attractive. Military aviation has all of the problems associated with civilian aviation; but on a more pressing schedule and requiring field maintenance. Tank and military automotive needs would also be pervasive.

SUMMARY/CONCLUSIONS

The Spline-Locking Screw presents a unique and fundamental building block to facilitate assembly and maintenance in space (micro "g"). Screw fastening , so pervasive on earth, could now be employed in micro "g" without danger of cross-threading by Astronauts in space suits or by robots. This would advance the capability of assembly, maintenance and materials handling in a fundamental sense. Further, by allowing standardization and modular construction on a here-to-for unprecedented scale, Spline-Locking Screws would simplify logistics; a consideration which is of special import in space.

The concept, while new, is straight forward. Indeed, Goddard Space Flight Center has already successfully constructed and demonstrated a Workpiece Attachment Mechanism/Workpiece Attachment Fixture (WAM/WAF) for the Flight Telerobotic Servicer (FTS) and End Effector Spline-Locking Screw prototypes for fastening payload boxes are in construction.
Commercial possibilities for this concept fall into an impressive market niche' , particularly for Bolts 0.5 in. dia. or larger. These include: 1. civilian aviation maintenance and overhaul, mechanical systems and avionics. 2. The computer industry. 3. Earth moving and materials handling systems. 4. The Automotive

industry. 5. Military applications in general with aviation, avionics and tank and automotive support in particular.

REFERENCES

1. ***Portion of Shop Drawing #GL 1087209, Rockwell International Corp. Space Division***, dated 9/7/76 showing the operation of an ACME **Screw Payload Fastener for the Solar Max repair mission of 1982**, courtesy of Robert Davis, NASA/GSFC.
2. ***FTS End Effector End Item Specification*** dated 1990, courtesy of Paul W. Richards, NASA/GSFC.
3. ***Invention Disclosure : SPLINE-LOCKING PAYLOAD FASTENER*** by John M. Vranish GSFC dated 8/10/90, Case no. GSC 13,378-1, filed 6/5/91.
4. ***Invention Disclosure: WORK ATTACHMENT MECHANISM/WORK ATTACHMENT FIXTURE (WAM/WAF)*** by John M. Vranish GSFC dated 2/25/91, Case no. GSC 13,430-1.
5. ***Invention Disclosure: SPLINE SCREW AUTOCHANGER*** by John M. Vranish GSFC dated 3/26/91, Case no. GSC 13,435-1.
6. ***Invention Disclosure : SPLINE-LOCKING PAYLOAD FASTENING SYSTEM*** by John M. Vranish GSFC dated 8/10/90, Case no. GSC 13,454-1.
7. ***Invention Disclosure: SPLINE SCREW MULTIPLE ROTATIONS MECHANISM*** by John M. Vranish GSFC dated 6/3/91, Case no. GSC 13,452-1.

N92-22691

QUICK APPLICATION/RELEASE NUT WITH ENGAGEMENT INDICATOR (COMMERCIAL APPLICATION OF AN INNOVATIVE NUT DESIGN)

Jay M. Wright
NASA Johnson Space Center
Houston, TX 77058

ABSTRACT

This is an assembly which permits a fastener to be inserted or removed from either side with an indicator of fastener engagement. The nut has a plurality of segments, preferably at least three segments, which are internally threaded, spring-loaded apart by an internal spring, and has detents on opposite sides which force the nut segments into operative engagement with a threaded member when pushed in and release the segments for quick insertion or removal of the fastener when moved out. When the nut is installed, end pressure on the detents presses the nut segments into operative engagement with a threaded member where continued rotation locks the structure together with the detents depressed to indicate positive locking engagement of the nut. On removal, counterclockwise rotation relieves the endwise pressure on the detents permitting internal springs to force the detents outward and allowing the nut segments to move outward and separate to permit quick removal of the fastener.

INTRODUCTION

Conventional nuts for fastening objects together have the disadvantage of requiring a large number of turns to position them in a fully locked position. This may also involve the application of a considerable amount of torque. Mechanical operations in space, i.e., at low gravity, encounter problems which do not exist on earth. For example, when one applies torque, as in tightening a threaded nut or joint, a countertorque is encountered against the worker (Newton's third law) which tends to rotate the worker around the object being torqued.

It is the object of this nut to provide a system that allows quick insertion and removal of a fastener system in those cases where time and cost is a major factor. Such as, oil and chemical applications where the thread form may be corroded or contaminated so that removal of a nut is time consuming and potentially costly.

The nut incorporates a positive locking system, which provides for guaranteed thread engagement, at specified depth of thread, for cases where thread engagement is critical to continued operations.

The nut incorporates for positive lock indication which provides for a visual check that the part is fastened properly.

The nut can be used freely or mounted to an object.

The fastener or nut may be inserted and torqued with the torque wrench, saving time and effort to first run the fastener down and then changing wrenches, or, if the torque wrench is used to run down the fastener, it saves on wear of the expensive torque wrench.

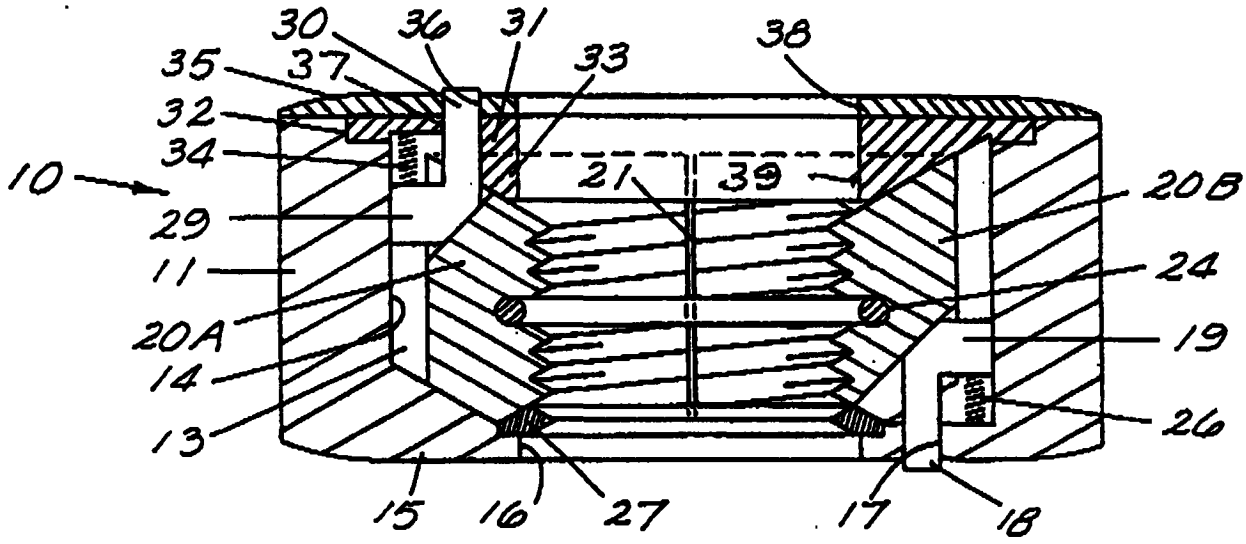


FIG. 1 is a sectional view, on the section line 1 - 1 of FIG. 2, of a quick application/release nut illustrating a preferred embodiment of this invention.

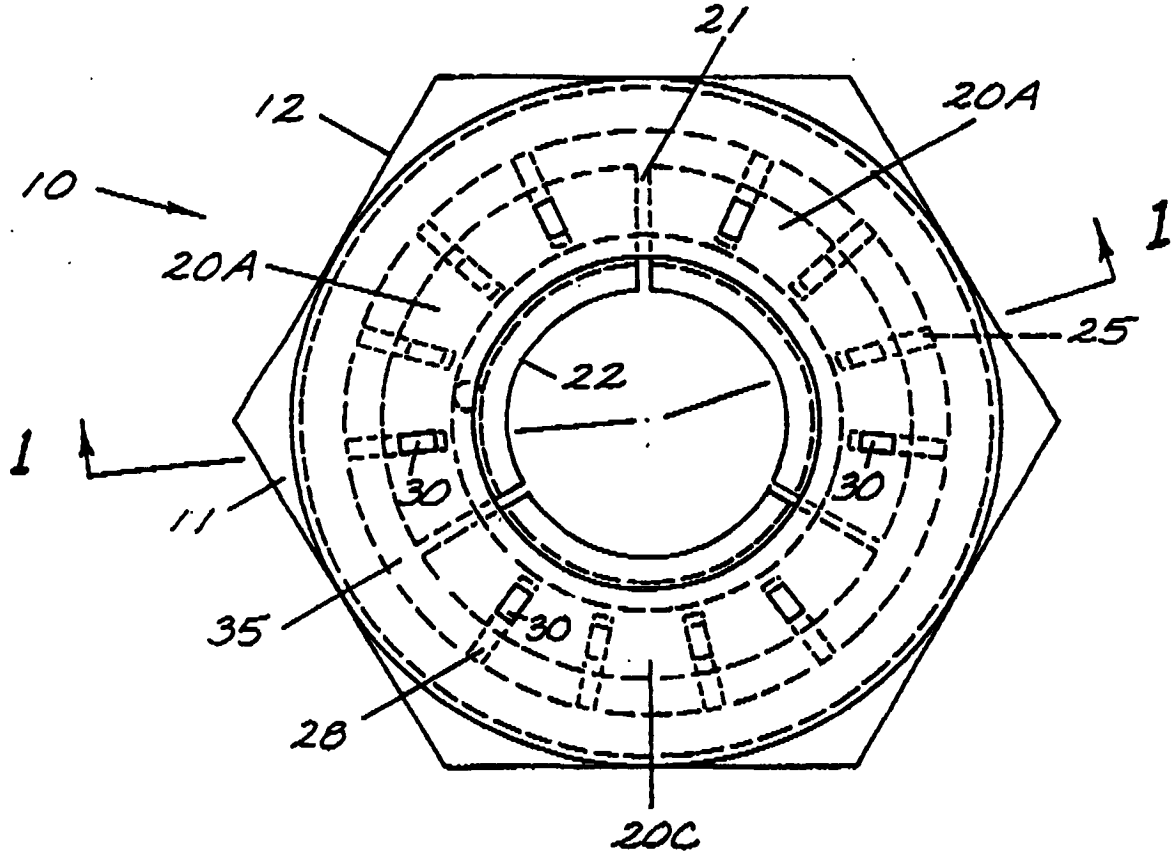


FIG. 2 is a top plan view of the nut shown in FIG. 1.

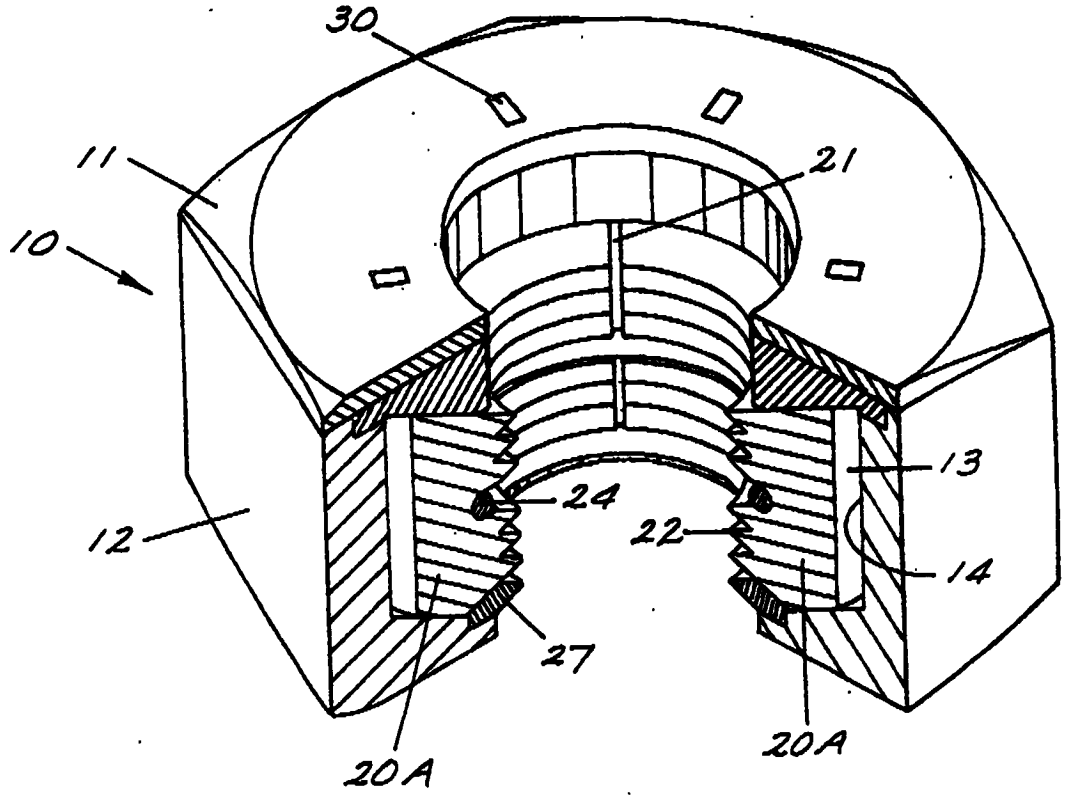


FIG. 3 is a front isometric view in partial section of the nut shown in FIG. 1.

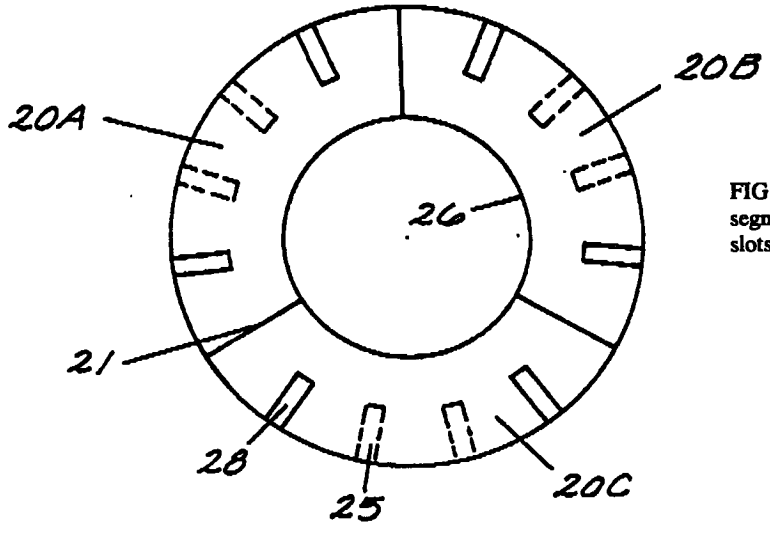


FIG. 4 is a top plan view of the multiple internally threaded segments used in the nut shown in FIG. 2.

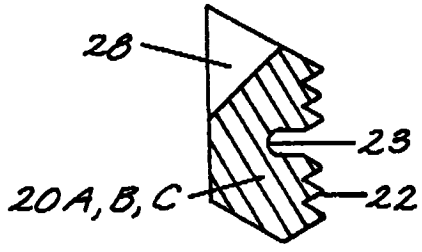


FIG. 6 is a sectional view of one of the internally threaded segments used in the nut shown in FIG. 1 showing the bottom slots for the operating detents.

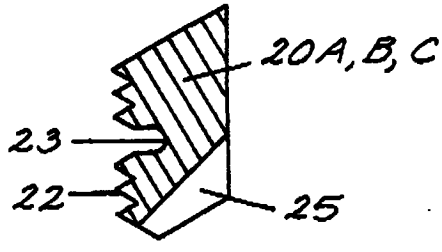


FIG. 5 is a sectional view of one of the internally threaded segments used in the nut shown in FIG. 1 showing the top slots for the operating detents.

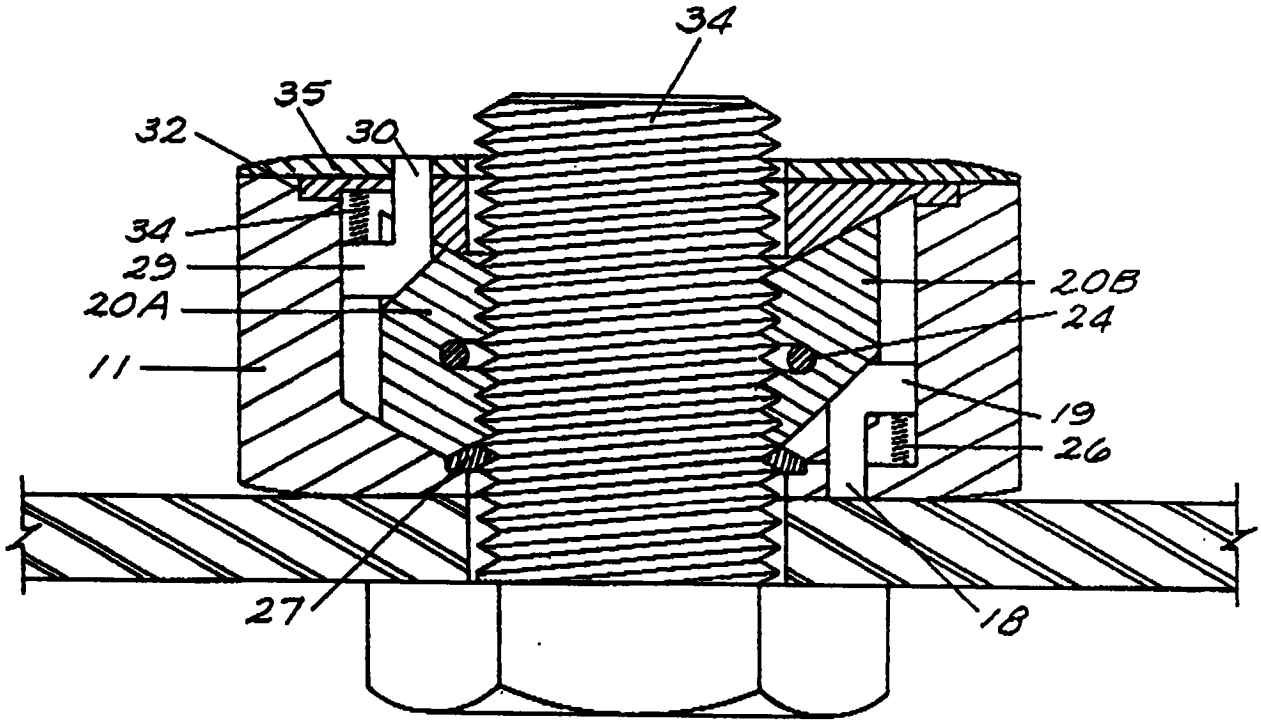


FIG. 7 is a sectional view, in longitudinal central section with a bolt or threaded shaft or spindle inserted from one side of the nut.

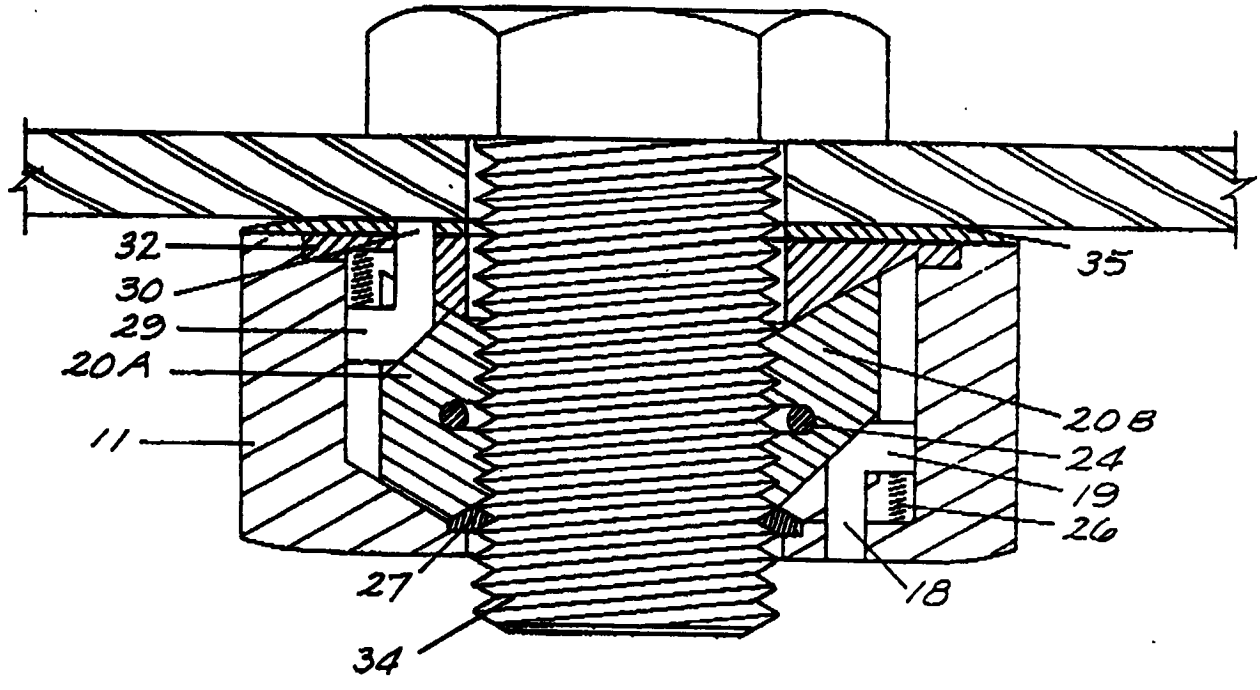


FIG. 8 is a sectional view, in longitudinal central section with a bolt or threaded shaft or a spindle inserted from another side of the nut.

N92-22692

INFLATABLE TRAVERSING PROBE SEAL

Paul A. Trimarchi
NASA Lewis Research Center
M.S. 86/12
Cleveland, OH 44135

ABSTRACT

An inflatable seal acts as a pressure-tight zipper to provide traversing capability for instrumentation rakes & probes. A specially designed probe segment with a teardrop cross-section in the vicinity of the inflatable seal minimizes leakage at the interface. The probe is able to travel through a lengthwise slot in a pressure vessel or wind tunnel section, while still maintaining pressure integrity. The design uses two commercially available inflatable seals, opposing each other, to cover the probe slot in a wind tunnel wall. Proof-of-Concept tests were conducted at vessel pressures up to 30 psig, with seals inflated to 50 psig, showing no measurable leakage along the seal's length or around the probe teardrop cross-section. This seal concept can replace the existing technology of sliding face-plate / O-ring systems in applications where lengthwise space is limited.

INTRODUCTION

This seal was conceived for a pressurized wind tunnel into which sensor probes are to be inserted and positioned from the outside. With the proposed seal, a probe can penetrate and move along a wind tunnel wall without causing an air leak at the wall. The probe can also be moved in and out through the seal, if desired.

A pair of opposed inflatable rubber seals (or tubes) form the airtight seal across a slot in a wind tunnel wall. The probe which penetrates this seal has a special double-teardrop cross section which allows the inflatable seal to deform and mold itself around the probe. (See Fig. 1). The probe can be traversed from one end of the seal to the other while differential pressure is maintained across the wall of the wind tunnel.

Current technology for commercial sealed traversing actuators utilizes sliding plates with face seals. Unfortunately, this method is very bulky, and the actuator/seal assembly must be at least twice as long as the desired probe travel (in order to keep the probe slot completely covered even when the probe is at the extreme ends of the slot). However, the new Inflatable Traversing Probe Seal assembly takes up only half as much length as commercially available actuators. Seal lengths only slightly longer than the desired probe travel are possible. This size advantage can make a tremendous difference when space is cramped, and will allow the probe to travel very close to an obstruction or flange.

This work was done by NASA-Lewis Research Center Engineering Directorate for the Aerodynamics Branch in support of the Supersonic Shear Flow Research Rig. A prototype seal test apparatus was constructed and tested, and preliminary design of a new spoolpiece and inflatable seal was completed in 1990. Tests proved that the concept will create an airtight seal while still allowing probe movement. Fabrication of actual wind-tunnel hardware has been postponed, however, due to a temporary lack of program funds to build the proposed spoolpiece for the Shear Flow Rig.

TESTING

The prototype model consisted of a 4" x 4" x 22" long square tube "pressure vessel" with an 18" long opposed inflatable seal assembly attached to one face. (See Fig. 2). The seal assembly was assembled from commercially available inflatable seals and extruded aluminum seal retainers. Both the square pressure vessel and the inflatable seal were pressurized with shop air through two separate regulators. The prototype

seal was tested at a vessel pressure of 30 psi, with the seal inflated to 50 psi. No leakage past the seal was measured, either with the probe stationary or in motion.

Force measurements were also recorded when the probe was in motion. The actuation force necessary to drive the probe through the inflatable seal was provided by mounting the teardrop-shaped dummy probe to the movable arm of a Tensile Test Machine. (See Fig. 3). Probe actuation forces of approx. 75 lbs. were measured with the seal lubricated with silicone O-ring grease. After the probe was stroked the length of the seal several times, actuation forces began to rise rapidly. Forces of up to 130 lbs. were measured after two strokes. The rise in actuation force was attributed to the silicone grease being wiped free after several strokes of the probe, causing a significant increase in friction.

Seal lubrication is required with such frequency that an active lubrication system was devised to bleed grease from the teardrop-shaped portion of the probe. Pressurized grease is fed from a reservoir, through a solenoid valve, down to the seal/probe interface. Tiny bleed holes on the teardrop-shaped probe section weep grease into the interface. The solenoid valve is controlled on demand by a load-sensing system which causes grease to bleed whenever the actuation force exceeds 100 lbs. The grease ceases to bleed once the actuation force drops below 80 lbs.

Another interesting observation noted during prototype testing was that the actuation force required to move the probe did not vary in proportion to the seal inflation pressure. Since the seal inflation pressure creates a frictional load on the teardrop-shaped dummy probe, it was assumed that probe actuation forces would rise with higher seal inflation pressures. However, testing indicated that probe actuation force did not vary measurably with seal inflation pressure. Actuation force was more dependent on the amount of lubrication remaining on the seal interface.

End termination method is an important aspect of this inflatable seal. The prototype utilized the Presray "Inflato-Boot™ " seal end configuration. (See Fig. 4). A specially made plexiglas wedge was mounted on either end of the seal holder to fill in the gap and give the boot something to seal against. (The wedge is visible in Fig. 3). Unfortunately, it was very difficult to achieve a leak-free seal junction using this method. Strategic injections of silicone rubber sealant during assembly finally solved the leakage problem. The Inflato-Boot™ is a unique option which is probably better suited to more conventional applications of inflatable seals. For NASA's application, it now appears that a standard end configuration utilizing a solid non-expanding portion, cast to the inflated configuration (rather than the usual relaxed configuration) would be easier to seal and less likely to leak. (See Fig. 5). It would also be less complicated, requiring no custom-machined wedge to bear against. This second method (standard end, inflated configuration) is the one which has been baselined for the NASA Shear Flow Research Rig application.

CONCLUDING REMARKS

This seal concept utilizes commercially available components to provide a dramatic improvement to the capabilities of sealed actuator systems. The concept is adaptable to other situations in which objects must penetrate pressure walls and move along them. It should work for vacuum chambers as well as pressure vessels. The concept may also be used for dust seals, regardless of differential pressure. A U.S. patent has been approved for the new device, with release scheduled for early 1992.

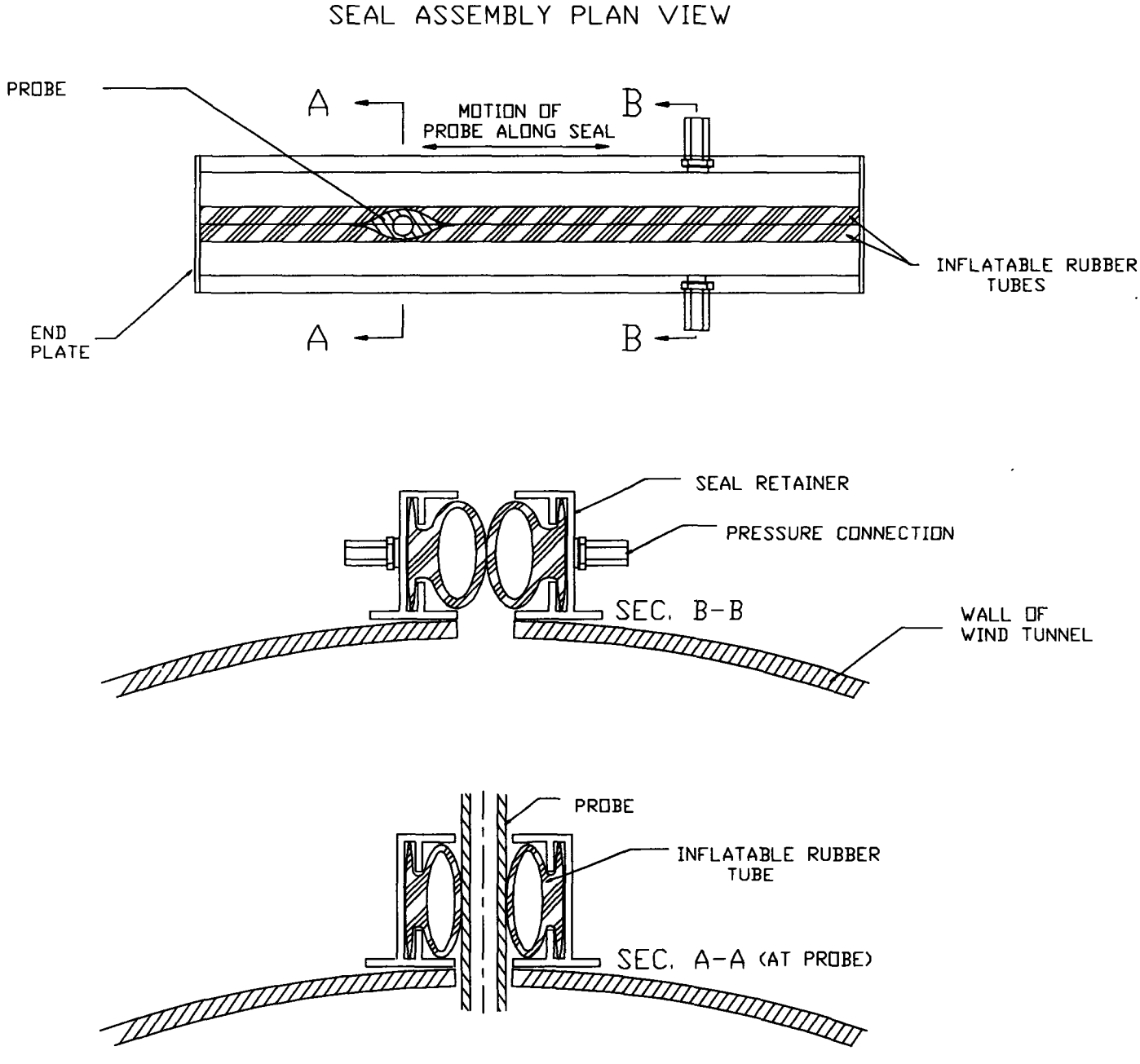

SEAL ASSEMBLY PLAN VIEW
PROBE
A
MOTION OF
PROBE ALONG SEAL
B
INFLATABLE RUBBER
TUBES
END
PLATE
A
B
SEAL RETAINER
PRESSURE CONNECTION
SEC. B-B
WALL OF
WIND TUNNEL
PROBE
INFLATABLE RUBBER
TUBE
SEC. A-A (AT PROBE)

FIGURE 1

NASA
C-89-13804

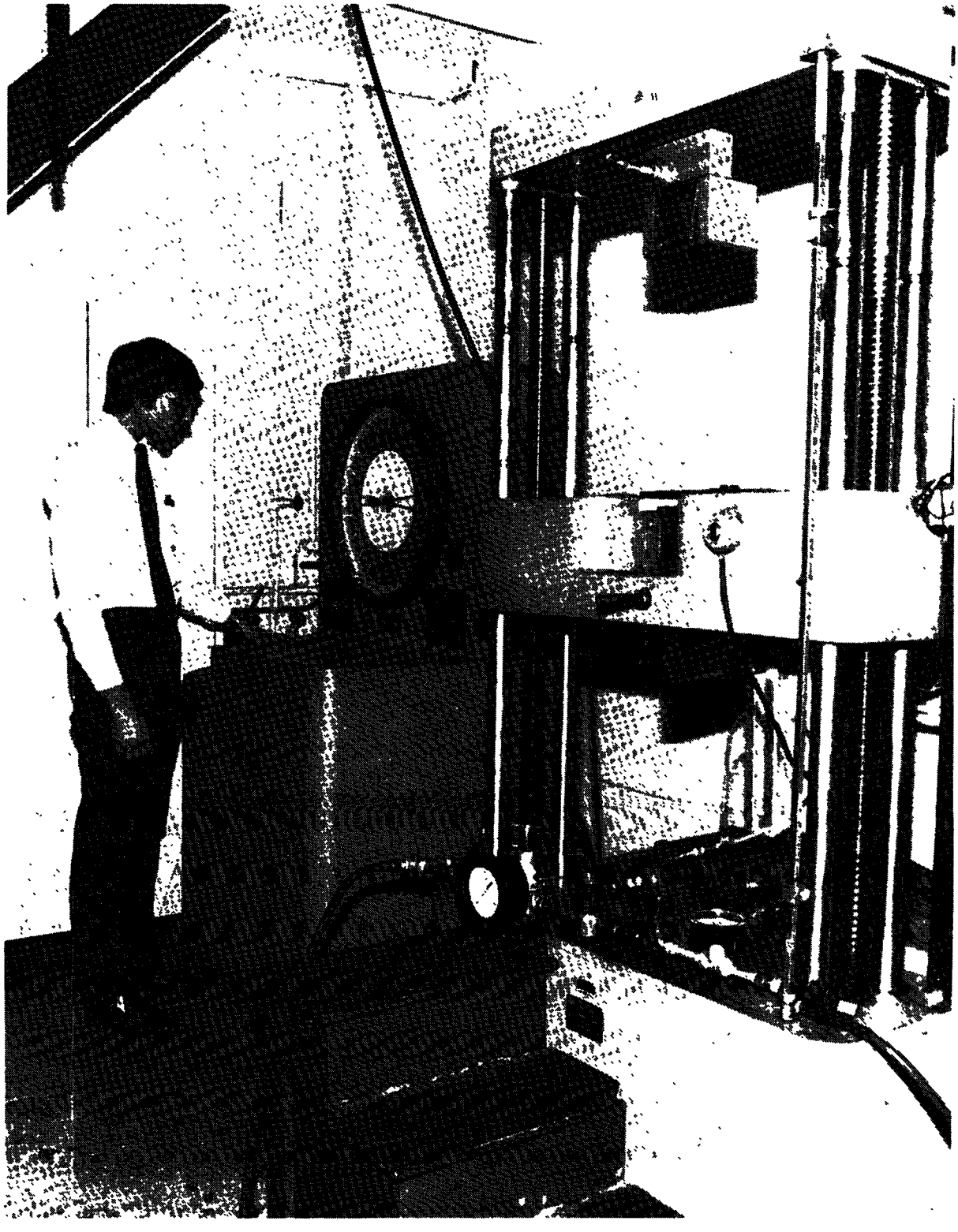


Fig. 2

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

NASA
C-89-13809

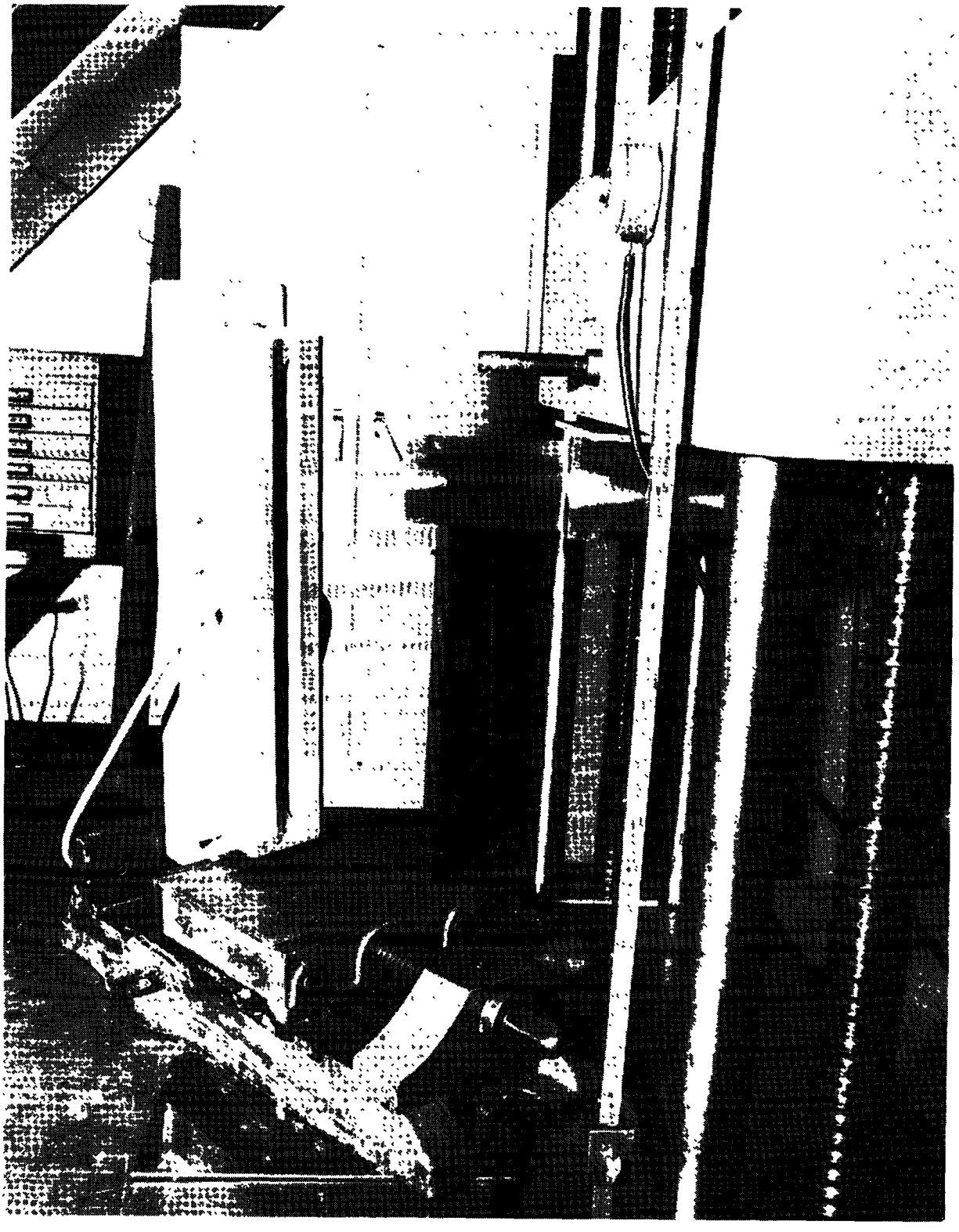


Fig. 3

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

PRESRAY'S NEW

INFLATO-BOOT™

Presray offers, as a standard feature in PRS 702, PRS 703, PRS 706, PRS 582, PRS 583 and PRS 705 seals, and as an optional feature in PRS 701 and PRS 580, a unique end configuration which minimizes the non-sealed area, and eliminates the need for end clamps:

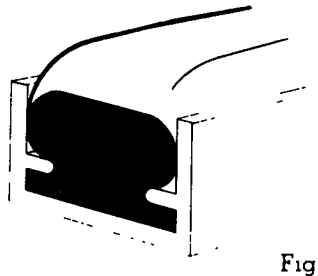


Fig. B

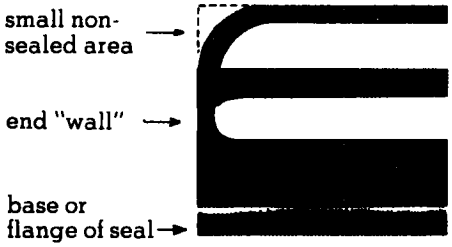
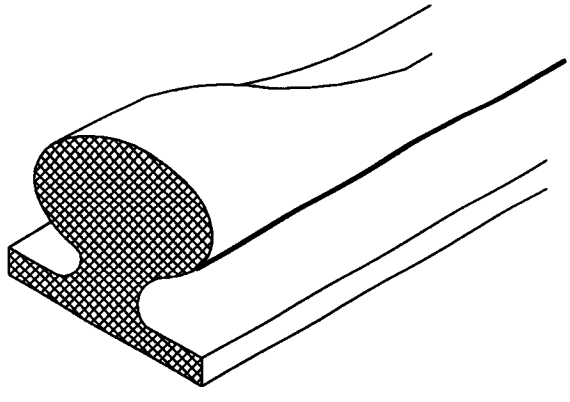


Fig. A
Side View

Patent Pending

Courtesy of Presray Corp.

Fig. 4



Standard Seal End Cast
Into Inflated Configuration

Fig. 5

ARTIFICIAL INTELLIGENCE

(Session D2/Room B1)

Thursday December 5, 1991

- **CLIPS: An Expert System Building Tool**
- **Fuzzy Logic Applications to Expert Systems and Control**
- **Neural Network Technologies**
- **From Biological Neural Networks to Thinking Machines**

 PRECEDING PAGE BLANK NOT FILMED

N92-22693

CLIPS: AN EXPERT SYSTEM BUILDING TOOL

Gary Riley
Software Technology Branch
NASA Johnson Space Center
Mail Stop PT4
Houston, TX 77058

ABSTRACT

Expert systems are computer programs which emulate human expertise in well defined problem domains. The potential payoff from expert systems is high: valuable expertise can be captured and preserved, repetitive and/or mundane tasks requiring human expertise can be automated, and uniformity can be applied in decision making processes. The C Language Integrated Production System (CLIPS) is an expert system building tool, developed at the Johnson Space Center, which provides a complete environment for the development and delivery of rule and/or object based expert systems. CLIPS was specifically designed to provide a low cost option for developing and deploying expert system applications across a wide range of hardware platforms. The commercial potential of CLIPS is vast. Currently, CLIPS is being used by over 3,300 individuals throughout the public and private sector. Because the CLIPS source code is readily available, numerous groups have used CLIPS as the basis for their own expert system tools. To date, three commercially available tools have been derived from CLIPS. In general, the development of CLIPS has helped to improve the ability to deliver expert system technology throughout the public and private sectors for a wide range of applications and diverse computing environments.

INTRODUCTION

Conventional programming languages, such as FORTRAN and C, are designed and optimized for the procedural manipulation of data (such as numbers and arrays). Humans, however, often solve complex problems using very abstract, symbolic approaches which are not well suited for implementation in conventional languages. Although abstract information can be modeled in these languages, considerable programming effort is required to transform the information to a format usable with procedural programming paradigms.

One of the results of research in the area of artificial intelligence has been the development of techniques which allow the modeling of information at higher levels of abstraction. These techniques are embodied in languages or tools which allow programs to be built that closely resemble human logic in their implementation and are therefore easier to develop and maintain. These programs, which emulate human expertise in well defined problem domains, are called expert systems. The availability of expert system tools has greatly reduced the effort and cost involved in developing an expert system.

The C Language Integrated Production System (CLIPS) [1, 2] is an expert system tool developed by the Software Technology Branch at NASA's Johnson Space Center. The prototype of CLIPS, version 1.0, was developed in the spring of 1985 in a UNIX environment. Subsequent development of CLIPS greatly improved its portability, performance, and functionality. The first release of CLIPS, version 3.0, was in July of 1986. The latest version of CLIPS, version 5.1, was released in October of 1991. A version of CLIPS written entirely in Ada, CLIPS/Ada, has also been developed. CLIPS is currently available to the general public through the Computer Software Management and Information Center (see appendix).

KEY FEATURES OF CLIPS

CLIPS was designed to address several issues key to NASA. Among these were the ability to run on a wide variety of conventional hardware platforms, the ability to be integrated with and embedded within conventional software systems, and the ability to provide low cost options for the development and delivery of expert systems.

CLIPS is written in C for portability and speed and has been installed on many different computers without changes to the source code. At the time of its original development, CLIPS was one of the few tools that was written in C and capable of running on a wide variety of conventional platforms. CLIPS can be ported

PRECEDING PAGE BLANK NOT FILMED

to any system which has an ANSI compliant C compiler including personal computers (IBM PC compatibles, Macintosh, Amiga), workstations (Sun, Apollo, NeXT), minicomputers (VAX 11/780, HP9000-500), Mainframes (IBM/370), and supercomputers (CRAY).

Figure 1. CLIPS is Easily Ported From One Environment to Another

To maintain portability, CLIPS utilizes the concept of a portable kernel. The kernel represents a section of code which utilizes no machine dependent features (see Figure 2). The inference engine contains the key functionality of CLIPS and is used to execute an expert system. Access functions allow CLIPS to be embedded within other systems. This allows an expert system to be called as a subroutine (representing perhaps only one small part of a much larger program). In addition, information stored in CLIPS can be accessed and used by other programs. Integration protocols allow CLIPS to utilize programs written in other languages such as C, FORTRAN, and Ada. Integration guarantees that an expert system does not have to be relegated to performing tasks better left to conventional procedural languages. It also allows existing conventional code to be utilized. The CLIPS language can also be easily extended by a user through the use of the integration protocols.

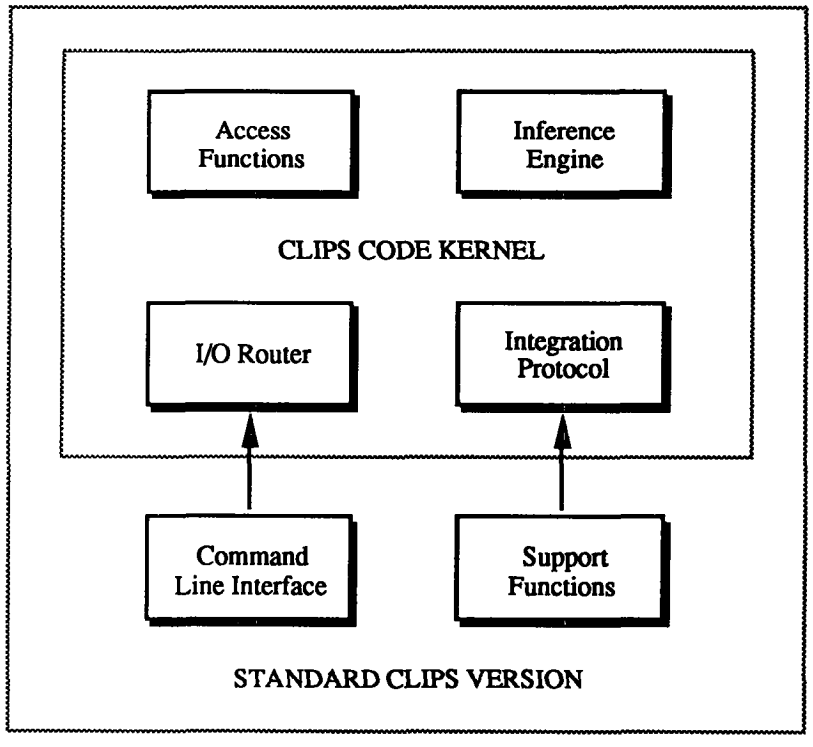


Figure 2. The CLIPS Code Kernel

To provide machine dependent features, such as windowed interfaces or graphics editors, CLIPS provides fully documented software hooks which allow machine dependent features to be integrated with the kernel. The I/O router system allows interfaces to be layered on top of CLIPS without making changes to the CLIPS kernel. The standard interface for CLIPS is a simple, text-oriented, command prompt. However, three interfaces are also provided with CLIPS that make use of the I/O router system and integration protocols to provide machine specific interfaces. These interfaces are provided for Apple Macintosh systems, IBM PC MS-DOS compatible systems, and X Window systems. Figure 3 shows the CLIPS interface for the Macintosh computer.

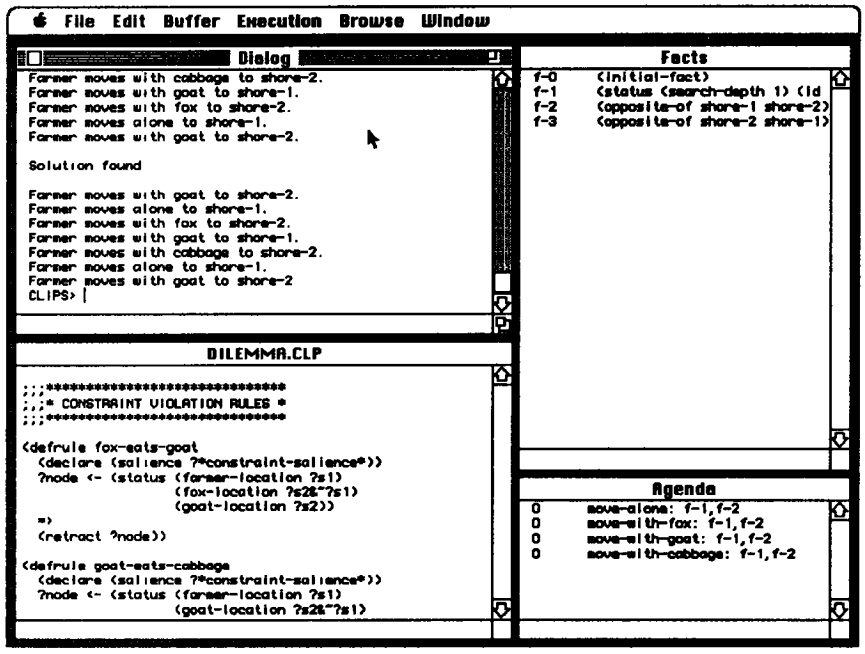


Figure 3. CLIPS Macintosh Interface

One of the key appeals of the CLIPS language results from the availability of the approximately 40,000 lines of CLIPS source code. Because the development of an expert system tool can require many man-years, the benefits of using CLIPS as a starting point for research and the creation of special purpose expert system tools cannot be understated. CLIPS users have enjoyed a great deal of success in adding their own language extensions to CLIPS due to the source code availability and its open architecture [3, 4, 5, 6, 7, 8]. Many users have also developed their own interfaces and interface extensions [9, 10, 11, 12].

KNOWLEDGE REPRESENTATION

Expert system tools are designed to provide highly productive environments by allowing knowledge to be represented flexibly. A flexible representation scheme allows the application developers to try several different approaches or to use an approach best suited to their problem. CLIPS provides a cohesive tool for handling a wide variety of knowledge with support for three different programming paradigms: rule-based, object-oriented, and procedural. In addition, CLIPS also supports the concepts of iterative refinement (refining an expert system with small iterative changes) and rapid prototyping (demonstrating proof of concept) which are found in many expert system tools.

Rule-Based Programming

The first (and originally the only) programming paradigm provided by CLIPS is rule-based programming. In this programming paradigm, rules are used to represent heuristics, or "rules of thumb", which specify a set of actions to be performed for a given situation. A rule is composed of an *if* portion and a *then* portion. The *if* portion of a rule is a series of patterns which specify the facts (or data) which cause the rule to be applicable. The process of matching facts to patterns is called pattern matching. CLIPS provides a mechanism, called the inference engine, which automatically matches facts against patterns and determines which rules are applicable. The *if* portion of a rule can actually be thought of as the *whenever* portion of a rule since pattern matching always occurs whenever changes are made to facts. The *then* portion of a rule is the set of actions to be executed when the rule is applicable. The actions of applicable rules are executed when the CLIPS inference engine is instructed to begin execution. The inference engine selects a rule and then the actions of the selected rule are executed (which may affect the list of applicable rules by adding or removing facts). The inference engine then selects another rule and executes its actions. This process, illustrated by Figure 4, continues until no applicable rules remain.

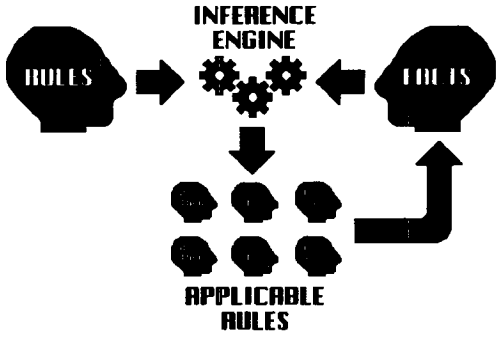


Figure 4. Execution of a Rule-Based Program

To illustrate the advantages of rule-based programming, consider the problem of monitoring a series of sensors. The following example program written in the C programming language illustrates how these sensors could be monitored using a procedural programming paradigm to determine if any two of the sensors have *bad* values (which a hypothetical expert indicates represents an overheated device).

```
#define BAD 0
#define GOOD 1
#define DEVICE_OVERHEATED 0
#define DEVICE_NORMAL 1

int CheckSensors(sensorValues,numberOfSensors)
  int sensorValues[];
  int numberOfSensors;
  {
   int firstSensor, secondSensor;

   for (firstSensor = 1;
        firstSensor <= numberOfSensors;
        firstSensor++)
      {
       for (secondSensor = 1;
            secondSensor <= numberOfSensors;
            secondSensor++)
          {
           if ((firstSensor != secondSensor) &&
               (sensorValues[firstSensor] == BAD) &&
               (sensorValues[secondSensor] == BAD))
             { return(DEVICE_OVERHEATED); }
          }
      }

    return(DEVICE_NORMAL);
   }
```

The *CheckSensors* function is implemented by storing the values of the sensors as integers in an array and then using two *for* loops to compare all combinations to determine if any two sensors have *bad* values. This function is relatively efficient if the sensors only need to be checked once. However, if this check is performed each time a sensor's value is changed, then all possible combinations are rechecked which is inefficient. In addition, the programmer has the responsibility for calling this function whenever an update is made to a sensor's value. An additional function could be written to check only one sensor against all other sensors, however, this increases the burden on the programmer. For contrast, the equivalent CLIPS code for a rule which performs the same task is shown following.

```
(defrule Two-Sensors-are-Bad
   (Sensor (ID-number ?id)   (status Bad))
   (Sensor (ID-number ~?id)  (status Bad))
    =>
   (assert (Device (status Overheated))))
```

The first line of the rule contains the keyword *defrule* which indicates that a rule is being defined. The symbol *Two-Sensors-are-Bad* is the name of the rule. The next two lines beginning with the symbol *Sensor* are the patterns that form the *if* portion of the rule. Essentially, the first pattern searches for any *Sensor* fact that contains a *status* value of *Bad* and the second pattern searches for another *Sensor* fact with a *status* value of *Bad* that does not have the same *ID-number* as the *Sensor* fact matching the first pattern. The => symbol serves to separate the *if* portion of the rule from the *then* portion of the rule. Finally, the *assert* command in the *then* portion of the rule creates a new fact which indicates that the device has overheated.

Because of the overhead associated with the inference engine and the generality provided through pattern matching, a rule-based program generally does not execute as quickly as a procedural program. However, significantly less code is required and the programmer does not have to explicitly check for applicable rules when sensor values are changed. Rules are always looking for new facts which satisfy their conditions. Indeed, careless implementation of pattern matching capabilities in a procedural language may result in a program which runs much less efficiently than its rule-based counterpart. CLIPS's inference engine is based on the Rete algorithm [13] which is an extremely efficient algorithm for pattern matching.

Object-Oriented Programming

The second programming paradigm provided by CLIPS is object-oriented programming. This programming paradigm allows complex systems to be modelled as modular components (which can be easily reused to model other systems or to create new components). Object-oriented programming encompasses a number of concepts including data abstraction (the ability to define complex objects using high level representations), encapsulation (the ability to hide the implementation details of an object, thereby increasing its modularity and potential for reuse), inheritance (the ability to define new classes of objects by reusing existing classes), and polymorphism (the ability of different objects to respond to the same "command" in specialized ways).

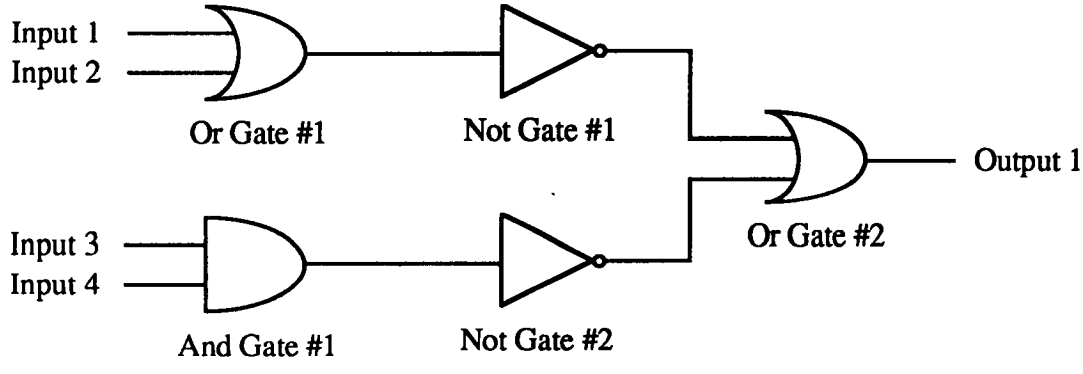


Figure 5. A Simple Electronic Circuit

Figure 5 shows a diagram of an electronic circuit consisting of *and*, *or*, and *not* gates. In electronics, a gate is a circuit that has an output dependent on some function of its input. The gates shown in Figure 5 all have boolean inputs and boolean output values. Physically, these boolean values would correspond to high and low voltages. Conceptually, these boolean values could be considered as *On* and *Off* or *True* and *False*. An *and* gate has an output value of *True* if all of its inputs are *True*, otherwise its output value is *False*. An *or* gate has an output value of *True* if any of its inputs are *True*, otherwise its output value is *False*. A *not* gate has an output value of *True* if its input value is *False* and an output value of *False* if its input value is *True*. In Figure 5, if *Input 1* and *Input 2* are both *False* and *Input 3* and *Input 4* are both *True*, then the output of *Or Gate #1* would be *False* and the output of *And Gate #1* would be *True*. The output of *Not Gate #1* would be *True* since its input (the output of *Or Gate #1*) is *False*. The output of *Not Gate #2* would be

False since its input (the output of *And Gate #1*) is *True*. Finally, *Output 1* from *Or Gate #2* would be *True* since at least one of its inputs (the output from *Not Gate #1*) is *True*.

Using object-oriented programming methodologies, it is relatively easy to model the behavior of the electronic circuit shown in Figure 5. The first step in modelling the circuit is to define classes which can be used to describe the gates used in the circuit. Since all of the gates might have some attributes in common (such as a part number), it would be useful to first define a *Gate* class. Another class, *One Input*, could be used to describe the attributes associated with a single input gate (such as a *not* gate). Since a two input gate is essentially a one input gate with an additional input, the *Two Input* class could inherit the attributes of the *One Input* class and then define additional attributes for the second input. Similarly, a *One Output* class and *Two Output* class could also be defined. Figure 6 illustrates the basic classes used to describe the gate circuits in Figure 5. The classes described illustrate the basic concepts of data abstraction and inheritance. Note that even though the circuit gates shown in Figure 5 would not need to utilize the *Two Output* class, other types of gates could utilize this class. For example, a *splitter* gate (which splits its one input into two identical outputs) could make use of this class.

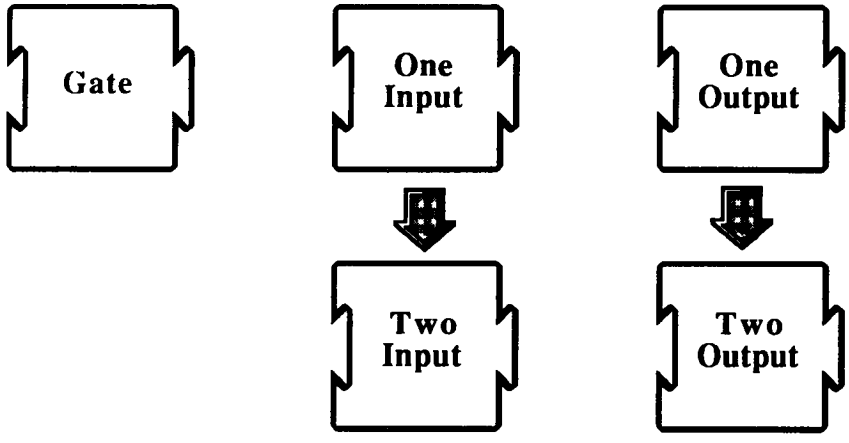


Figure 6. Classes Used to Describe Electronic Circuit Gates

Once the base classes for the gates are defined, it is possible to describe the gates in terms of these classes. Figure 7 conceptually illustrates how this could be done for the *not* gate and the *and* gate. The type of inheritance shown in Figure 7 is called multiple inheritance since a single class is inheriting attributes from more than one class. For example, the *And Gate* class inherits attributes from the *Two Input*, *One Output*, and *Gate* classes. In contrast, the inheritance shown in Figure 6 is called single inheritance since a single class inherits attributes from at most one other class (such as the *Two Input* class inheriting attributes from the *One Input* class). Some object-oriented programming languages support only single inheritance. CLIPS provides support for full multiple inheritance.

The following code shows how the gate classes could actually be defined in the CLIPS programming language (assuming the base classes described previously have been defined).

```
(defclass Not-Gate
  (is-a One-Input One-Output Gate))

(defclass And-Gate
  (is-a Two-Input One-Output Gate))

(defclass Or-Gate
  (is-a Two-Input One-Output Gate))
```

Each class definition contains the keyword *defclass* which indicates that a class is being defined. This keyword is followed by the name of the class. The next line of each class definition indicates the classes from which the class being defined will inherit attributes. Inheritance is specified using the *is-a* keyword. If

desired, additional attributes or slots of a class can be defined after the inheritance is specified. For example, the *Two Input* class might be defined as shown in the following code.

```
(defclass Two-Input
  (is-a One-Input)
  (slot Input-2))
```

Once the gate classes have been defined, it is possible to define instances (or objects) of these classes. For example, *Or Gate #1* would be a specific instance of the *Or-Gate* class as would *Or Gate #2*. It would have its own data areas for storing its input and output values. Thus a class serves as the prototypical definition which is used for creating objects belonging to that class.

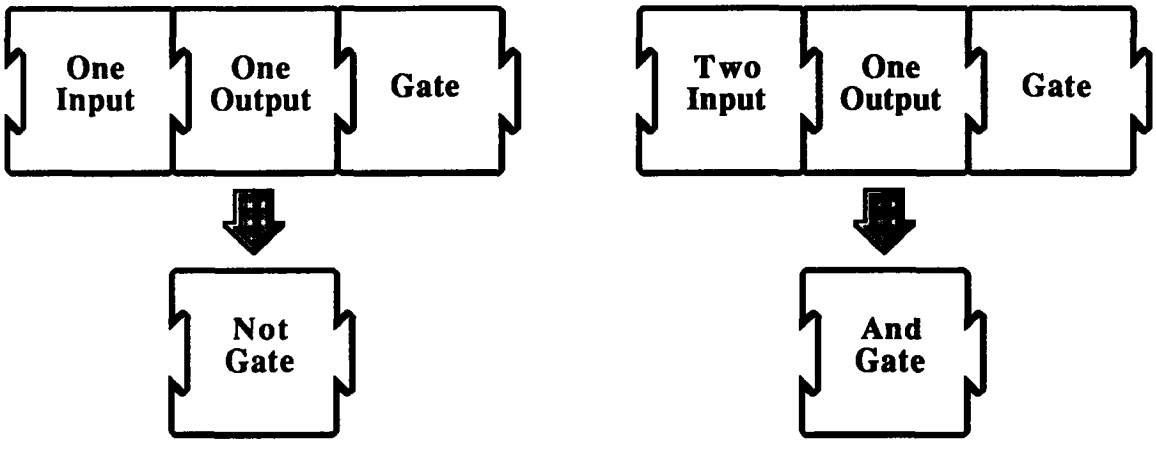


Figure 7. Building New Classes from Existing Classes

In CLIPS, objects are manipulated by sending them messages which specify an action to perform. For the circuit example, an appropriate action might be to recompute the output of a gate based upon its inputs. Notice that even though the *or* gates and *and* gates are both *Two-Input One-Output Gates*, their outputs are computed differently. In object-oriented programming, procedures as well as data can be associated with objects. Rather than writing one routine to compute the output values for all gate types given their inputs, the routines for computing outputs for objects can be encapsulated inside the classes themselves. When an *or* gate is sent a *Compute Output* message, its output is computed to be *True* if either of its inputs are *True*, otherwise its output is *False*. When an *and* gate is sent a *Compute Output* message, its output is computed to be *True* if both of its inputs are *True*, otherwise its output is *False*. Thus, both objects respond differently, yet appropriately, to the same message. This behavior is the essence of polymorphism and is illustrated by Figure 8. The procedures attached to classes are referred to as message-handlers. The following CLIPS code shows how message-handlers could be defined for the gate classes.

```
(defmessage-handler Not-Gate Compute-Output ()
   (put Output-1 (not (get Input-1))))

(defmessage-handler And-Gate Compute-Output ()
   (put Output-1 (and (get Input-1) (get Input-2))))

(defmessage-handler Not-Gate Compute-Output ()
   (put Output-1 (or (get Input-1) (get Input-2))))
```

Each message-handler definition contains the keyword *defmessage-handler* which indicates that a message-handler is being defined. This keyword is followed by the name of the class for which the message-handler is being defined and then the name of the message handled by the message-handler. The next line of each definition contains the single action performed by these message-handler. In general, message-handlers can perform as many actions as are required to complete their task. The task of computing the output requires only one action. The *put* function is used to change the value of an object's attribute.

For each of these message-handlers, the attribute being changed is the *Output-1* attribute. The value to which this attribute is changed varies for each message-handler, but is based on the inputs of the object. The *get* function is used in the message-handlers to retrieve the values of the *Input-1* and *Input-2* attributes. The message-handlers for the *Not-Gate*, *And-Gate*, and *Or-Gate* classes use the *not*, *and*, and *or* functions respectively to compute the correct output value based on their inputs. When writing CLIPS code, a prefix notation is used for calling functions that is very similar to the LISP programming language (even though CLIPS is written in the C programming language).

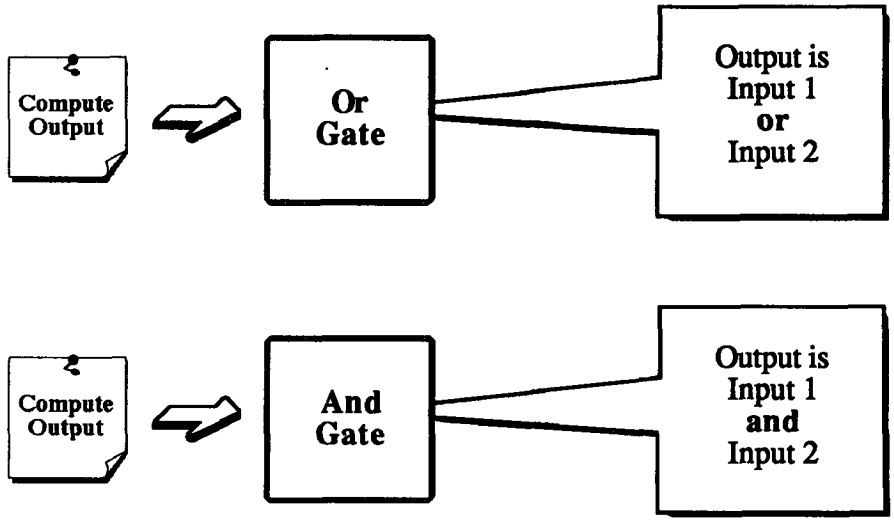


Figure 8. Two Different Objects Responding Differently to the Same Message

Procedural Programming

The third programming paradigm provided by CLIPS is procedural programming. This programming paradigm provides capabilities similar to those found in languages such as C, Pascal, Ada, and LISP. With respect to building expert systems, these are the least interesting capabilities provided by CLIPS. However, the ability to define procedural code directly within CLIPS allows new procedural capabilities to be added to CLIPS without the need of a compiler or linker. To add new capabilities to CLIPS which have been written in languages such as C, FORTRAN, or Ada, a compiler and linker are required to recompile and relink the new source code with the CLIPS source code. CLIPS allows the definition of global variables, functions, and generic functions. Generic functions are the most interesting feature of the CLIPS procedural programming language in that they allow different pieces of procedural code to be executed depending upon the arguments used when calling a function. This capability is called function overloading. As an example, the addition function could be overloaded so that numeric data types are numerically added and string data types are concatenated.

CURRENT USES

Although CLIPS was originally developed to aid in the construction of aerospace related expert systems, it has been put to widespread usage in a number of fields. CLIPS is being used by over 3,300 users throughout the public and private community including: all NASA sites and branches of the military, numerous federal bureaus, government contractors, 170 universities, and many companies. At the First and Second CLIPS Conferences held in August 1990 and September 1991 respectively, over 120 papers were presented on a diverse range of topics. In addition to aerospace and engineering applications, some other examples of CLIPS applications include: software engineering [14], network security [15], genetics [16], medicine [17], botany [18], and agriculture [19]. To date, three commercially available tools have been derived from CLIPS.

CONCLUSION

Because of its portability, extensibility, capabilities, and low-cost, CLIPS has received widespread acceptance throughout the government, industry, and academia. The development of CLIPS has helped to improve the ability to deliver expert system technology throughout the public and private sectors for a wide range of applications and diverse computing environments.

REFERENCES

1. *CLIPS Reference Manual*, Version 5.1, NASA document JSC-25012, Houston, TX., September 1991.

2. Giarratano, J., and Riley, G. *Expert Systems: Principles and Programming*, Boston, PWS-KENT, 1989.

3. Snyder, J., and Chirica, L. "An SQL Generator for CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

4. Bhatnagar, H., Krolak, P., McGee, B., and Coleman, J. "A Neural Network Simulation Package in CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

5. Orchard, R., and Diaz, A. "BB_CLIPS: Blackboard Extensions to CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

6. Homeier, P., and Le, T. "ECLIPS: An Extended CLIPS for Backward Chaining and Goal-Directed Reasoning," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

7. Adler, R. "Integrating CLIPS Applications into Heterogeneous Distributed Systems," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

8. Boyle, C., and Schuette, J. "Debugging Expert Systems using a Dynamically Created Hypertext Network," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

9. Pickering, B., and Hill, R. "HyperCLIPS: A HyperCard Interface to CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

10. Callegari, A. "Integrating Commercial Off-the-shelf (COTS) Graphics and Extended Memory Packages with CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

11. Jenkins, J., Holbrook, R., Shewhart, M., Crouse, J., and Yarost, S. "CLIPS Application User Interface for the PC," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

12. Feagin, T. "On the Generation of Graphical Objects and Images From Within CLIPS Using XView," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

13. Forgy, C. "Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem," Pages 17-37, Artificial Intelligence 19, (1982).

14. Morris, K. "Automating Symbolic Analysis with CLIPS," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

15. Miller, M., Barr, S., Gryphon, C., Keegan, J., Kniker, C., and Krolak, P. "The Management and Security Expert (MASE)," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

16. Inglehart, J., and Nelson, P. "LinkFinder: An Expert System That Constructs Phylogenic Trees," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

17. Salzman, G., Duque, R., Braylan, R., and Stewart, C. "A CLIPS Expert System for Clinical Flow Cytometry Data Analysis," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

18. Heymans, B., Onema, J., and Kuti, J. "Application of a Rule-Based Knowledge System Using CLIPS for the Taxonomy of Selected *Opuntia* Species," Proceedings of the Second CLIPS Conference, Houston, TX., September 1991.

19. Engel, B., Jones, D., Rhykerd, R., Rhykerd, L., Rhykerd Jr., C., and Rhykerd, C. "A CLIPS Expert System for Maximizing Alfalfa (Medicago Sativa L.) Production," Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

APPENDIX

CLIPS is free to NASA, USAF, and their contractors for use on NASA and USAF projects by calling the Software Technology Branch Help Desk between the hours of 9:00 AM to 4:00 PM (CST) Monday through Friday at (713) 280-2233. Government contractors should have their contract monitor call the Software Technology Branch Help desk to obtain CLIPS. Others may obtain CLIPS through the Computer Software Management and Information Center (COSMIC), which is the distribution point for NASA software. The program number is COS-10022. The program price is $350.00, and the documentation price is $140.00 (as of August 1991). The program price is for the source code. Price discounts are available to U.S. academic institutions. Further information can be obtained from

COSMIC
382 E. Broad St.
Athens, GA 30602
(404) 542-3265

An electronic bulletin board containing information regarding CLIPS can be reached 24 hours a day at (713) 280-3896 or (713) 280-3892. Communications information is 300, 1200, or 2400 baud, no parity, 8 data bits, and 1 stop bit.

FUZZY LOGIC APPLICATIONS TO EXPERT SYSTEMS AND CONTROL

N92-22694

Dr. Robert N. Lea
Software Technology Branch/PT4
NASA Johnson Space Center
Houston, TX 77058
Phone: 713/483-8085
Email: rlea@nasamail.nasa.gov

Yashvant Jani, Ph.D.
Technology Systems Division
Togai InfraLogic Inc.
Houston, TX 77058
Phone: 713/480-8904
Fax: 713/480-8906

ABSTRACT

A considerable amount of work on the development of fuzzy logic algorithms and application to space related control problems has been done at the JSC over the last few years. Particularly, guidance control systems for space vehicles during proximity operations, learning systems utilizing neural networks, control of data processing during rendezvous navigation, collision avoidance algorithms, camera tracking controllers, and tether controllers have been developed utilizing fuzzy logic technology. These systems have given very good results, and in some areas such as fuel and power usage they have shown superior performance to shuttle systems. Several other areas in which fuzzy sets and related concepts are being considered at the JSC are to diagnostic systems, control of robotic arms, pattern recognition and image processing.

It has become evident, based on the commercial applications of fuzzy technology in Japan and China during the last few years, that this technology should be exploited by the government as well as private industry for energy savings, reducing human involvement in industrial processes, and for many complex control problems where precise mathematical modeling is either practically impossible or very costly.

1. INTRODUCTION

In his article "Fuzzy Sets", Prof. Zadeh [1] first developed the concepts of fuzzy logic in 1965 and established these concepts firmly during the 70's [2] with other pioneers [3,4,5,6,7]. Contrary to its name, it is a precise subdiscipline in mathematics that enables mathematicians and engineers to utilize human like thinking in decision making processes. Handling imprecise information is much easier in the architecture provided by fuzzy logic compared to conventional logic. However, it was not accepted by the U.S. community for a long time simply because the word 'fuzzy' has negative connotations. It was felt that fuzzy logic based decision making processes involved fuzzy reasoning rather than human like common sense reasoning that must sometimes be based on information that is inherently not crisp. In fact, fuzzy logic is a method based on sound mathematical principles that enables one to model natural language rules and to make common sense evaluations of the degree to which non-crisp conditions such as low temperature, fast speeds, or sharp turns are satisfied. Control engineers made arguments that desired control can be achieved using existing control theory principles until the triple inverted pendulum was balanced using fuzzy logic principles. This task could not be performed adequately using conventional logic. The real utility of this logic was shown by the Japanese [8,9,10] when they applied the principles to subway control, automatic transmission control, camera focusing, and many other problems. Recently U.S. business people have taken an interest in this field [11,12,13] as the Japanese applications have shown that this logic can save developmental time and costs.

Investigations in the areas of fuzzy logic and neural networks have been underway since 1984-1985 [14] in the Software Technology Laboratory (STL) at the Johnson Space Center (JSC). The utility of this logic in several autonomous space operations applications have been demonstrated utilizing high fidelity orbital operations simulations [15]. Our objectives in the STL are to investigate new technologies for control and decision making processes. We are evaluating fuzzy logic techniques, neural network methods, genetic algorithms, and learning techniques based on fuzzy sets and neural networks for building expert systems and robust controllers. We are

particularly investigating the feasibility of applying these technologies to space operations to achieve desired operational efficiency and reduce overall life cycle costs.

In this paper, we describe the fuzzy logic applications achieved at JSC in section 2. Our current activities in the areas of fuzzy learning systems and neural networks are described and potential results are discussed in section 3. Several commercial applications mainly from Japan are described in section 4 and a summary and discussion of advantages and disadvantages of fuzzy logic is given in section 5.

2. FUZZY LOGIC APPLICATIONS AT JSC

As reported earlier [14,15], several fuzzy logic and neural networks applications are underway at JSC. In this section we will summarize results of only three because of space limitations: the six degree-of-freedom (6DOF) controller, the collision avoidance, and the camera tracking control.

2.1 Proximity Operations And Results

The 6 DOF controller for a spacecraft has been designed and tested [16,17] in a shuttle simulation for proximity operations. The 6 DOF controller uses sensor measurements of range, elevation and azimuth angle directly as input and generates the commands for the jet select routine to null out the errors. For a given mission profile, it maintains a proper range and range rate. The elevation and azimuth angle measurements are used in conjunction with the angular rates to generate jet firing commands required to follow a desired trajectory such as shown in fig. 1. For example, during the v-bar approach, the controller maintains elevation and azimuth angles close to zero and range and range rate close to their desired values. If the range is smaller than the desired range, the controller will slow down accordingly. If the range rate is slower than the desired rate then the controller will increase the speed. In keeping the elevation and azimuth angles close to zero, the controller adjusts its actions based on the angle errors, rate errors as well as current pitch rate and roll rate. While the vehicle is translating its attitude is being maintained by the rotational part of the controller.

Our approach of correcting the elevation and azimuth errors and range rate error in conjunction with maintaining orientation and body rate errors has given good results and shown significant savings in fuel (Table I). The performance of our 6 DOF controller based on simultaneous relative trajectory and attitude control is very good and robust. The controller is responsive and maintains the flight profiles within the expected range. The controller holds proper elevation and azimuth angles during all proximity operations test cases, and performs proper range and range rate control. It transitions along the v-var or r-bar from approach to station keeping in a way very consistent with profiles flown by pilots. It also performs fly-around maneuvers very well and continuously maintains proper range deadband and attitude while transitioning to the required station keeping position.

We plan to continue to test this 6 DOF controller further and compare its performance with mission planning data, the manual crew procedure test cases flown in mission simulator, and possibly flight data. Our preliminary test data shows that the correlation between the translational and rotational rates can be handled easily by the fuzzy controller. We also plan to modify the 6 DOF controller such a way that it can be easily adapted by other spacecraft.

2.2 Collision Avoidance

Future unmanned missions to Mars will investigate the terrain and collect soil samples in advance of manned missions [18]. Path planning is a crucial element in the activities to be undertaken by an autonomous rover. As an initial effort to address this problem a fuzzy control system for maneuvering a four wheel vehicle with front wheel steering from one position to another and requiring a particular attitude at the terminal point was developed in the STL [19,20]. Since obstacles such as boulders or troughs may block the shortest path from the current position to the target position for the next sample acquisition, collision avoidance algorithms were later developed [21,22] that takes sensor data giving range to obstacle, current velocity and orientation of the vehicle and processes it through fuzzy rules to generate steering commands to avoid obstacles as they are encountered. Simulation testing has been performed for a set of representative test cases, and performance of the guidance algorithms have been evaluated for a variety of obstacle scenarios. It was found that a higher-level path planner is needed to handle situations when the vehicle is caught in a back-off setting, that is, when it is not possible for the vehicle to pursue a "forward" path. It is significant to note that the method employed does not depend on object identification, but rather, detection of the degree to which an object (where present) or (more generally) an angular sector represents an obstacle. This is a significant relaxation over most collision avoidance schemes. Our simulation results and planned enhancements for

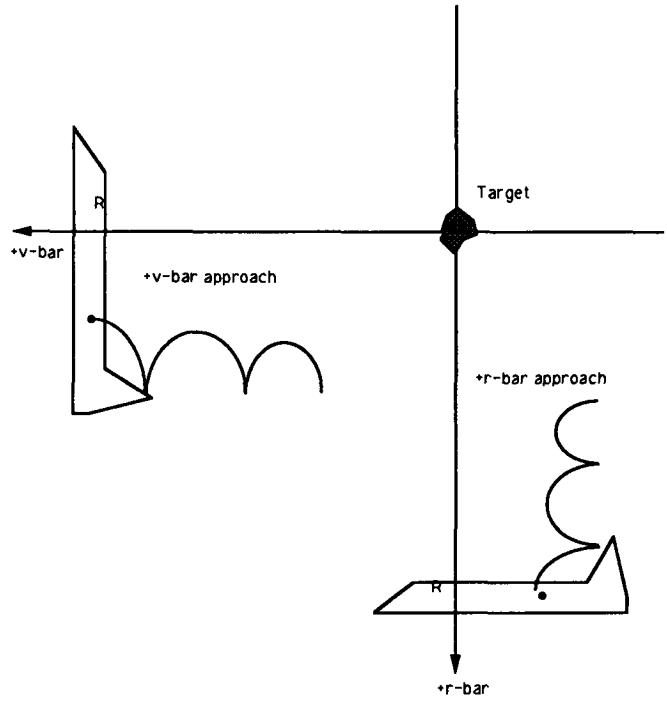


Fig. 1a Proximity Operations v-bar and r-bar Approaches

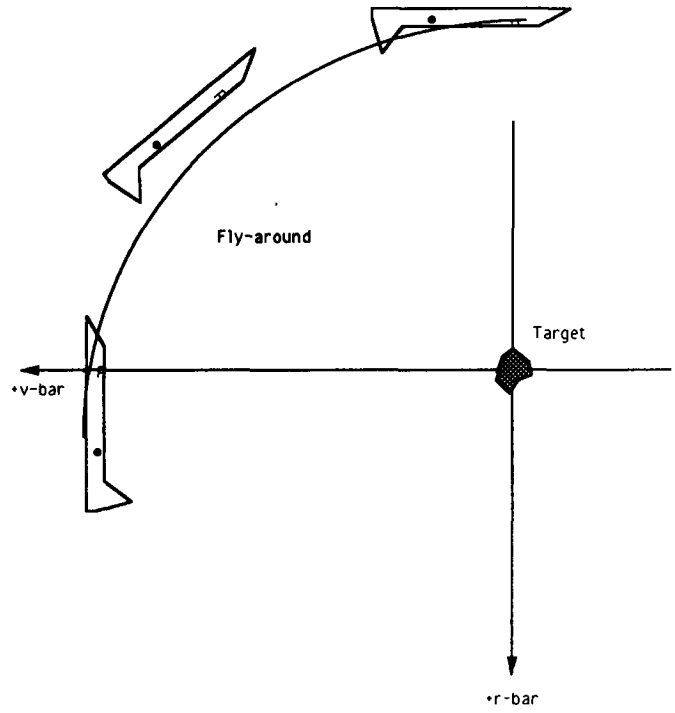


Fig. 1b Proximity Operations fly-around Segment

Table Ia . Fuel Usage Comparison between conventional and fuzzy attitude controllers

Run #	Fuel Usage (kg)		Savings
	Conventional	Fuzzy	%
5	18.40	18.44	-0.22
6	8.88	9.45	-6.42
7	26.27	20.58	21.66
8	159.91	122.26	23.54
13	24.22	23.31	3.76

Table Ib. fuel usage, arrival times and mission segment information for the 6 DOF testcases

RUN #	Maneuver	Distance	Rates	Arrival Time	Fuel Used
1	V bar approach	400 ft-50 ft	zero rates	1450 sec	33 lbs
2	R bar approach	400 ft-50 ft	zero rates	1720 sec	61 lbs
3	1/4 Fly Around	@200 ft	0.2 rate	1800 sec	48 lbs
4	Station Keeping	@200 ft	zero rates	1800 sec	12 lbs

the future point to refinements in the algorithm, the possibility of adaptive tuning of the system and, as expected, the need for a higher-level path planner to handle cases that involve backoff, sensor fusion, positioning of the vehicle at the destination, moving obstacles, and other situations that involve radically changing environments of operation.

Five test cases were designed to test the capabilities of the collision avoidance system. The first and second test cases were designed to test avoidance of a single obstacle with varying size. The third and fourth test cases show that obstacles can be successfully avoided one by one as the rover reaches the desired destination. The fifth test case shows that the rover can avoid many obstacles even if they are lined up, leaving only a small opening. The five test cases described above are given in Fig. 2, which was taken from the summary piece for each run from a graphical simulation developed on an IRIS workstation.

Some important points of these results are as follows: a) trajectory control [19,20] includes not only position control but also orientation control. When collision avoidance algorithms are integrated with the trajectory control [21,22], the orientation control can not be managed properly without major modification to the system. Proper orientation can be achieved only if there is sufficient distance between the last obstacle avoided and the target point so that the rover can turn itself. Since this distance can not be guaranteed, the desired orientation of the rover at arrival was not addressed with the collision avoidance problem; b) the back-off situation requires some knowledge about the obstacles just avoided, or some information about the obstacles in the path. Since our cameras were looking forward, we postponed the 'back-off' situation study; and, c) when the rover is going forward, and it encounters an obstacle that can be avoided only by going back, it must remember that the distance it must go back is a factor. Otherwise, it can get into a situation where it continues to go back and forth when the critical distance for collision avoidance is not altered. Thus, the transition from forward to backward requires special care.

Fuzzy logic control together with straightforward algorithms yield an effective system for autonomous collision avoidance in an environment of uncertain information. The technique is robust and avoids complexity in initial stages where simple obstacle avoidance is a key element rather than involved object identification or mapping of a complete world model. This technique could be integrated with fast and sophisticated object identification algorithms if desired. Future developments would include a plan for backoff situations where a forward path is not possible, fusion of information from several sensors of various types with differing degrees of uncertainty to be managed, adaptive tuning of fuzzy components, avoidance of moving obstacles, and a higher level path planner for optimization of operation in regional or global environments.

2.3 Camera Tracking Controllers

The concept of a camera tracking controller that generates the necessary pan and tilt motors commands to keep an object as close to the center of the field of view of the camera as possible has been developed at the JSC [23,24]. This concept has been developed further and currently software simulations have been built that are allowing limited performance testing. This Camera Tracking System (CTS) is a good example of one where a fuzzy control system is much better suited to the problem than conventional control methods. Clearly since one does not know ahead of time what object will be tracked or what it will do once the camera is trained on it. The fuzzy controller generates rate commands based on the required angle change, available camera rates, and the estimate of range to the object.

Seven test cases have been designed to provide representative trajectories during proximity operations like the V-bar approach, R-bar approach, station-keeping and fly around for the CTS performance evaluation. Since the object in our simulation does not have active control, its trajectory is purely based on the forces of the orbital environment. To give an example of results, the plots of the sixth test case which represents a Passing Orbit are discussed. In this test, the object is translating in all three direction as shown in fig. 3a, 3b, and 3c, where the LVLH x, y, and z position are given as functions of time. The out-of-plane and in-plane LVLH trajectories are shown in fig. 3h, and 3i. Because of this motion, the controller must use pan and tilt angles to track the object. The pixel-x, pan angle, pixel-y and tilt angle are shown in fig. 3d, 3e, 3f, and 3g respectively.

Initially, the camera is pointed along the LVLH x-axis. The pixel-x and pixel-y measurements are 240 and 280 respectively due to the position of the object in the LVLH frame. Thus, the controller commands pan left and tilt down. Pan left implies negative pan angle and therefore pan angle continuously decreases from a small negative value to a large one as shown in fig. 3e. The trajectory of the object is such that the pan angle continuously increases in the left direction making a full circle. In fig. 3e, the pan angle is nearly 2π at the end of the test. This behavior is completely expected for the Passing Orbit case.

Fig. 2a Collision Avoidance for an object

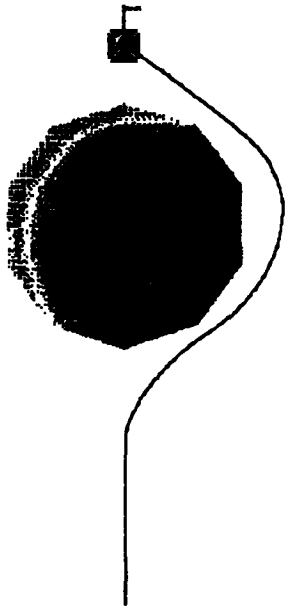


Fig. 2b Collision Avoidance for a large object

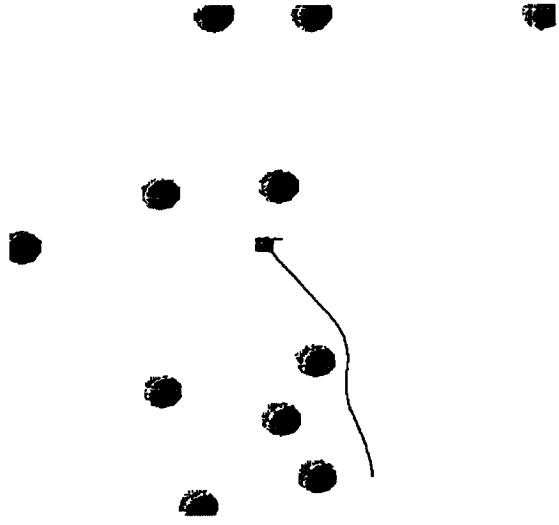


Fig. 2c Collision Avoidance for a set of objects (easy)

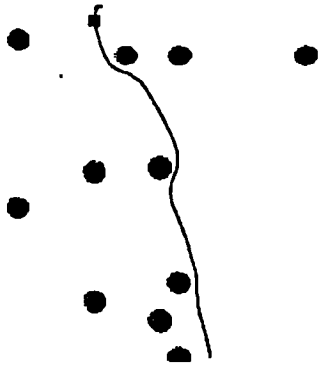


Fig. 2d Collision Aviodance for a set of objects (hard)

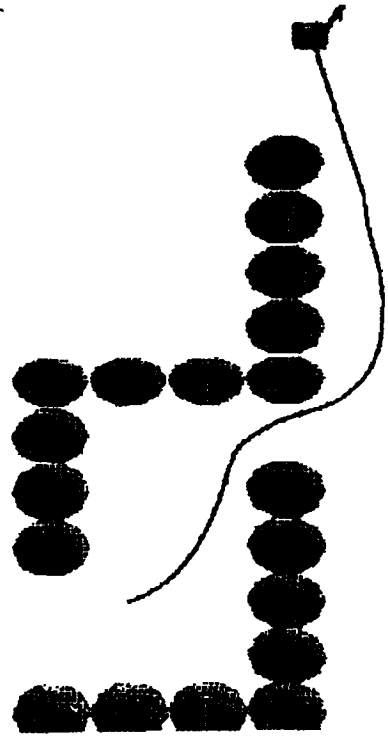


Fig. 2e Collision Avoidance for a simple maze

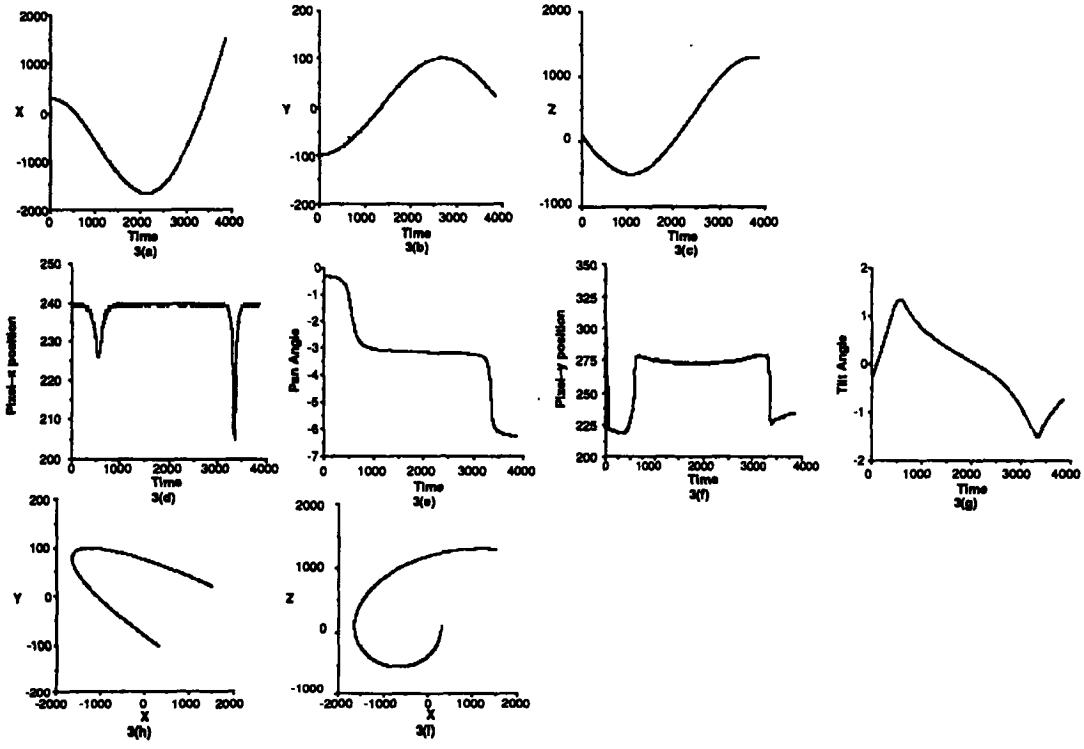


Fig. 3(a)-3(c) motion of the object in terms of X, Y, and Z vs. time in the LVLH coordinate system.
Fig. 3(d) correlates with 3(e), as the pixel–x position changes, the camera will pan right or left.
Fig. 3(f) correlates with 3(g), as the pixel–y position changes, the camera will tilt up or down.
Fig. 3(h) motion of the object in the X–Y plane in the LVLH coordinate system.
Fig.3(i) motion of the object in the X–Z plane in the LVLH coordinate system.

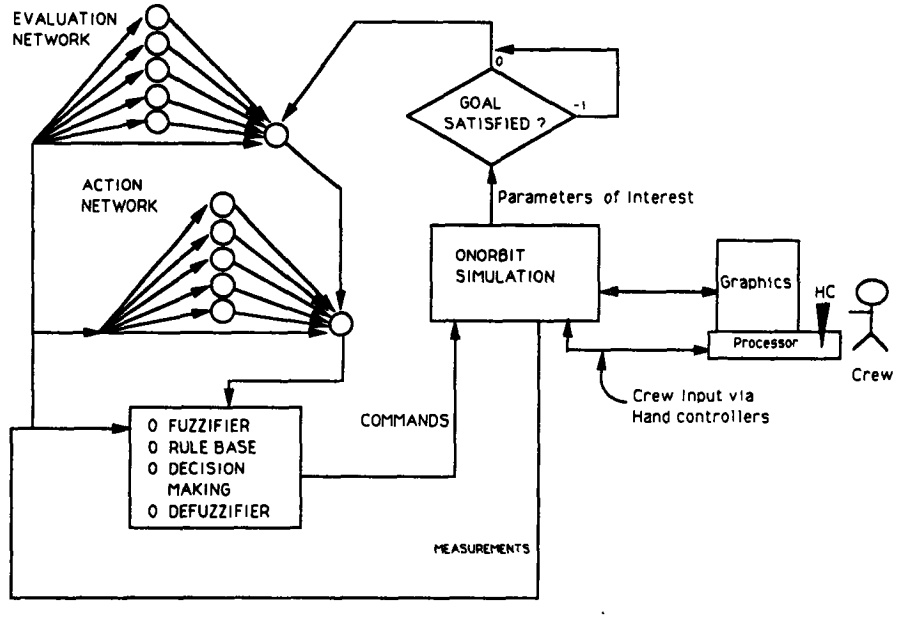


Fig. 4 FUZZY LEARNING SYSTEM FOR DOCKING OPERATIONS

The Tilt down command corresponds to a positive rate in the gimble frame, but results in a total negative tilt angle with respect to LVLH due to the mounting of gimble drive. The tilt angle decreases for only a couple of seconds. (Since the output data is every 10 seconds, tilt down is not visible in the plot.) As the pixel-y position drops below 270, the tilt rate command stops. The controller begins commanding tilt up, when the pixel-y position has a value less than 240. As the controller commands tilt up rate, the tilt angle increases in the positive direction as shown in fig. 3g. (Tilt up is negative rate in gimble frame, but results in positive rate with respect to LVLH frame and hence the positive tilt angle.) The tilt angle does not change during the time pixel-y position reverses from 224 to 275. However, when pixel-y is pegged at 275, the tilt angle changes from 1.5 to -1.5 radians.

In summary our test results show that : 1) the x-position and y-position of the object in the camera FOV are maintained in the zero-zone (pixel range 240 to 270) most of the time, confirming that the controller is capable of keeping the object in the camera's FOV, 2) pan and tilt angles were adjusted to reposition the object in the center (see definition of CENTER in ref. 24) when it went outside of the zero zone, 3) in one case, response of the controller was slower in comparison to the object's motion, however, the object never went out of the range of the CENTER membership function. Performance of the fuzzy tracking controller is very good and the object is maintained properly in the center of the FOV.

Our future activities for the CTS are focused on several tasks : 1) Modify the membership functions to maintain the object in a narrow range, 2) Utilize the range directly into the rules rather than the intermediate scale factor, and if possible reduce the number of rules, 3) implement the estimated angular velocity rules and provide a controller with a dynamic response, 4) implement the centroid calculations algorithms to handle the image in the pixel form, and 5) implement the rule set in a fuzzy chip that can directly interface with the gimble drives and camera digitizer.

3. CURRENT ACTIVITIES

Reinforcement learning techniques based on fuzzy control and multi-layer neural networks have been successfully demonstrated at the Ames Research Center (ARC) using the inverted pendulum as an example [25,26]. As a joint project between JSC and ARC, a concept has been put together for applying this technique to spacecraft docking operations [27]. Fuzzy controllers developed at JSC will be implemented in the Approximate Reasoning Intelligent Control (ARIC) architecture, fig. 4, and test cases will be performed using high fidelity simulation of shuttle docking scenarios. As a first part of this project the attitude controller developed for the shuttle has been implemented in the ARIC architecture and a test case for attitude hold has been performed. Our preliminary results are very promising [28], and shows that the fuzzy learning technique has no problem in controlling the angular rates. We are further investigating how to control angles and plan to expand the control to translational parameters such as range and elevation angle. Failure criteria has been developed for the attitude hold, and the fuel usage criteria will be developed as we further learn to utilize this technique for all attitude maneuvers.

The Tethered Satellite System (TSS) mission [29] planned for August 1992 involves the upward deployment of the Italian satellite on a 20 km. conducting tether. Electric current will be induced as the tether is dragged through the magnetic field of the Earth. Because of interaction between the orbital environment and tether dynamics, there are several oscillations in the tethered system. Particularly of interest are the "skip rope" oscillations resulting from the interaction of the Earth's magnetic field with the current pulsing through the tether. Identification and control of the skip rope oscillations is very important. Mission success is dependent on how successfully these oscillations can be damped during the satellite retrieval. Because of this situation, the mission profile has a very special Onstation-2 phase where the tether is not retrieved unless the skip rope magnitude is smaller than a critical value. Since the skip rope oscillations excite very peculiar attitude oscillations for the satellite, the magnitude and phase of the skip rope can be detected using the satellite rates and angle data. The coupling between the attitude oscillations and the skip rope behavior is complex and very non-linear.

The Space Time Neural Network (STNN) developed at JSC [30] provides the capability to learn the time variations as well as spatial variations using digital filters before the data is provided to hidden layers as well as the output layer. A concept under consideration is to apply the STNN to learn the coupling between satellite attitude oscillations and the tether skip rope behavior. Input to the network will be satellite rates, tether length, tension and other relevant parameters. Output should be the skip rope magnitude and phase. Using a 2.4 km. Onstation-2 test case, simulation data has been generated that can be used for training as well as testing. Several STNN's have been configured to identify the skip rope magnitude and phase. Twenty and thirty hidden nodes in two different layers have been used, and ten, twenty, and forty zeros for the digital filters have been used between the input and hidden layers. Preliminary results suggest that the STNN may require a long learning time. This is consistent with the

understanding that the coupling between the skip rope and the satellite oscillation is very complex and non-linear. Additional data may need to be generated for training. Data will be generated using tests like retrieval from 20 to 2.4 km., and Onstation-1 segment.

Other problems being considered in the STL are the use of fuzzy robot arm positioning control to relieve the difficulty of solving the inverse kinematics problem, diagnostics systems to help identify faults in a timely manner, Rotary Fluid Management Device pump control to conserve energy aboard the Space Station Freedom, and Station re-boost thrust and fuel management. Projects to study the use of fuzzy logic to conserve energy through efficient control of heating, ventilation, and air conditioning systems will be started this next year. These types of applications will be beneficial to many industrial operations.

4. COMMERCIAL APPLICATIONS OF FUZZY LOGIC

Applications described in this section are primarily based on information received from Dr. Masaki Togai of Togai InfraLogic in his presentation at a workshop in Nov. 1990 [31].

The fuzzy autofocusing system developed by Cannon and Togai InfraLogic and the image stabilization system for the Panasonic video camera recorder have been successfully implemented into their respective products. The objectives of the autofocus were to improve the quality of focus and reduce the focusing time. The technique is based on generating the approximate measure of sharpness and utilizing it in connection with the motor speed using fuzzy rules. There are 13 simple rules such as " If the sharpness is high and its differential is low then the focus motor speed is low". Results were very satisfying. The quality of focus was improved in terms of sharpness, the focusing time was reduced by 20 %, and the hunting was reduced resulting in power savings and motor wear and tear. The fuzzy logic based code is smaller in size than standard focusing systems and execution time is less. The Panasonic image stabilization system successfully eliminates unwanted movement in video recordings such as that caused by the camera being bumped or bouncing caused by unstable platforms such as would result if one were filming from a moving vehicle.

The concept and first design for the Mitsubishi Heavy Air Conditioner system was conceived in April 1988. The fuzzy inverter air conditioner system is based on 50 fuzzy rules and uses the max-product inferencing method and centroid defuzzification method. A temperature sensor provides the measurement and commands are generated for the inverter, compressor valve and fan motor in terms of inverter frequency, compressor pressure and fan speed. Simulation was completed by the summer of 1988 and production began in October 1989. The results are very encouraging. Room heating and cooling times were reduced to one-fifth, temperature stability increased by a factor of 2, there was an overall power savings of 24%, and a reduced number of sensors was required for the entire operation of system.

The objective of the Nissan Automotive Transmission application was to provide a "smoother ride", reduce wear on the transmission and provide a more "human" like shifting pattern. This fuzzy transmission controller utilizing a rule base replaces conventional control in one of Nissan's models. Testing and evaluation is not complete but initial results show a distinct reduction in the frequency of shift in a varied terrain.

Developed by Bunji Kaneko, manager of systems development at Yamaichi Securities, and Michio Sugeno, professor at Tokyo Institute of Technology, the Yamaichi Fuzzy Fund is a premiere application for trading systems. It handles 65 industries and a majority of the stocks listed on Nikkei Dow and consists of approximately 800 fuzzy rules. Rules for the system are determined by a committee that meets monthly. Additional changes are made by senior analysts as deemed necessary. The three major categories of rules are Macro rules, Micro rules and Industrial rules. The system was tested for two years, and its performance exceeds the Nikkei Average by over 20 %. While in testing, the system recommended "sell" 18 days before Black Monday. Analysts will agree that the rules for trading are all "fuzzy".

5. SUMMARY

Fuzzy logic is simple, easy to understand and reflects human type thinking. Its architecture is very well suited for implementing heuristic knowledge or the knowledge gained through experience. For example, control of a processing plant typically performed by human operator can be easily automated using this framework in software. Several applications have shown that the fuzzy logic based control is usually robust, non-linear and comparatively stable. It

provides an ability to combine seemingly non-related parameters for higher order reasoning. Control of systems that are non-linear and difficult to model is easily achieved using fuzzy logic principles.

Hardware and Software tools are now available such that systems can be implemented from concept to software simulation and then to hardware prototype in a very short time. Tools like the TIL-shell [32] and fuzzy-C compilers [33] provide capabilities to enter the fuzzy algorithms in graphical forms and generate error free source code for simulation as well as hardware chip. In our experience, fuzzy controllers can be easily designed using heuristics and experiencial knowledge. Tuning of membership functions and modifications to the rule base is also simplified at a point where rapid implementation and testing is possible. Maintenance of the algorithms is minimal. Since the algorithms are in a graphical form, the knowledge transfer from one generation to another generation is very easy. Applications like Cannon Auto-focusing Systems have shown that the fuzzy algorithms require small memory and process the data in faster in comparison with other algorithms.

In U.S., the word "fuzzy" has a bad connotation, and therefore, the industry is afraid that if their appliances are based on fuzzy logic, nobody will buy or use them. The market share will be lost resulting in less profit. Even though fuzzy logic is based on the well defined notion of a fuzzy set, confusion still exists because the word "fuzzy" does not seem to fit with words like logic or focusing. Since fuzzy logic allows modelling of human like thinking, control of processes begin to look very easy, and thus the engineers feel that since much of the complexity is lost something of importance must be missing. However, it should be emphasized that to build fuzzy control systems that work efficiently a through engineering understanding of the problem is required. The advantage is that very complex and ill-defined plants do not need to be modeled. It is true that for fuzzy control criteria such as stability, controllability and observability are not yet clearly defined. Work in this field is in progress and the community expects new results in a short time. Control criterion such as the Lyapunov criterion have been developed for fuzzy control and have been applied to aircraft systems. The utility of such criteria is being investigated for other applications and decision making processes.

Based on the many successes of this technology, for the most part by the Japanese, it appears to be about time to exploit this field for decision making and expert system applications and enjoy advantages offered by this logic and its architecture for efficiency, and cost savings. In space operations, autonomy at a higher level will be easier to achieve using this technology and the result will be a major contribution to cost effective operational efficiency.

REFERENCES

1. Zadeh, L. : "Fuzzy Sets", Information and Control, vol. 8, pp. 338-353, 1965.
2. Zadeh, L. : "Fuzzy Algorithms", Information and Control, vol. 12, pp. 94-104, 1968.
3. Zadeh, L. : "Outline of a new approach to the analysis of complex systems and decision processes", IEEE Trans. Syst., Man and Cyberns., vol. SMC-3, p. 28-44, 1973.
4. Dubois, D. and Prade, H. : Fuzzy Sets and Systems--Theory and Applications, Academic Press, N.Y. 1980.
5. Zimmermann, H.J. : Fuzzy Set Theory -and Its Applications, Kluwar-Nijhoff Publishing, 1985.
6. Klir G. J. ; and Folger T. A. : Fuzzy sets, Uncertainty, and Information, Prentice Hall, New Jersey, 1988.
7. Kosko, B. : Neural Networks and Fuzzy Systems, Prentice Hall, Englewood Cliffs, New Jersey, 1991.
8. Yasunobu, S. ; and Miyamoto, S. : "Automatic Train Operation System by Predictive Control", Industrial Applications of Fuzzy Control, Sugeno, M. (Ed.), 1-18, North-Holland: Amsterdam, 1985.
9. Shingu, T. ; and Nishimori, E. : Fuzzy Based Automatic Focusing System for Compact Camera, Proceeding of IFSA89, pp. 436-439, 1989.
10. Tobi, T. ; Hanafusa, T. ; Itoh, S. ; and Kashiwagi, N. : Application of Fuzzy Control System to Coke Oven Gas Cooling Plant, Proceedings of IFSA89, pp. 16-22, 1989.
11. First Industrial Conference on Fuzzy Logic Systems sponsored by MCC, the consortium of Corporations, Austin, June 1990.
12. IEEE International Conference on Fuzzy Systems FUZZ-IEEE '92, sponsored by the IEEE Neural Network Council, San Diego, CA, March 1992.
13. First International Workshop on Industrial Applications of Fuzzy Control and Intelligent Systems sponsored by Texas A&M University, College Station, Texas, November 1991.
14. Lea, R.N., and Jani, Y.K. : "Applications Of Fuzzy Logic To Control And Decision Making", Proceedings of Technology 2000 Conference, NASA Conference Publication 3109, vol. 2, 1990.
15. Lea, R.N. and Jani, Y.K. : "Fuzzy Logic In Autonomous Orbital Operations", Proceedings of the second Joint Technology Workshop on Neural Networks and Fuzzy Logic, NASA Conf. Pub. 10061, vol. 2, p. 81, 1990.

16. Lea, R.N., Hoblit, J., and Jani, Y. : "Performance Comparison of a Fuzzy Logic based Attitude Controller with the Shuttle On-orbit Digital Auto Pilot", NAFIPS '91 Workshop Proceedings, pp 291-295, 1991.
17. Lea, R.N., Y. Jani and Jeff Hoblitt, "A Fuzzy Logic Based Spacecraft Controller for Six Degree of Freedom Control and Performance Results," Proceedings of the AIAA conference on Guidance, Navigation and Control, New Orleans, La., August 1991.
18. R. Kahl and S. Bailey, "Mars Rover Sample Return: Project Description and Mission Operations Review", NASA JSC (New Initiatives Office), Houston, TX., Presented at MRSR Phase A Midterm Review, Nov. 1989.
19. Lea, R. N.; Walters, L.; and Jani, Y. K. : A Fuzzy Logic Approach to Mars Rover Trajectory Planning and Control, Proceedings of the 1st ISMCR, Session D.3, pp. D.3.1.1, June 1990.
20. R. Lea, "Fuzzy Logic Approach to Mars Rover Guidance", presented to The International Conference on Fuzzy Logic and Neural Networks IIZUKA '90, Iizuka, Japan, July 1990.
21. R. Lea, M. Murphy, and L. Walters, "Fuzzy Logic Control for Autonomous Collision Avoidance", NAFIPS '91 Workshop Proceedings, University of Missouri-Columbia, May 1991.
22. Lea, R.N., Y. Jani, M. Murphy, M. Togai, "Design and Performance of a Fuzzy Logic Based Vehicle Controller for Autonomous Collision Avoidance," Proceedings of the Fuzzy and Neural Systems, and Vehicle Applications, Tokyo, Japan, November 8, 1991.
23. Lea, R. N., Giarratano, J., Fritz, R. H., and Jani, Y. K., Fuzzy Logic Control for Camera Tracking System, abstract only, Proceedings of the 8th International Congress of Cybernetics and Systems, New York, June 1990.
24. Lea, R. N., R. Fritz, Y. Jani, and J. Giarratano," Fuzzy Logic Control for a Camera Tracking System," Proceedings of the Workshop on Space Operations Automation and Robotics held at NASA/JSC, Houston, Texas., July 1991.
25. Lee, C. C. ; and Berenji, H. R. : An Intelligent Controller Based On Approximate Reasoning And Reinforcement Learning, Proceedings of IEEE International Symposium on Intelligent Control, Albeny, NY 1989.
26. Berenji, H.R. : "Strategy Learning in Fuzzy Logic Control", Proceeding of North American Fuzzy Information Processing Society Workshop, Columbia, Missourri, May 14-17, 1991
27. Lea, R. N., Jani, Y. K., and Berenji, H. : Fuzzy Logic Controller with Reinforcement Learning for Proximity Operations and Docking, Pres. at the 5th IEEE Int. Symposium on Intelligent Control, vol. 2, p. 903, 1990.
28. Lea, R.N., H. Berenji and Y. Jani, "Approximate Reasoning-Based Learning and Control for Proximity Operations and Docking in Space,"Proceedings of the AIAA conference on Guidance, Navigation and Control, New Orleans, La., August 1991.
29. Stefano Coledan : "Tethered Satellite Advances", Space News, vol. 2, no. 15, p. 8, 1991.
30. Villarreal, J.A., and Shelton, R.O. : "A Space-Time Neural Network", Proceedings of the Second Joint Technology Workshop on Neural Networks and Fuzzy Logic, NASA Conf. Pub. 10061, vol. II, P. 63, 1991
31. Togai, M. : Commercial Applications of Fuzzy Logic in Japan , Presentation at the workshop on Fuzzy Control Systems and Space Station Applications (sponsored by NASA/Ames Research Center and McDonnell Douglas Space Systems Company Space Station Division), Huntington Beach, California, November 1990.
32. Hill, G.; Horsthotte, E.; and Teichrow, J. : TIL Shell User's Manual, Togai InfraLogic Inc., Jan. 1990.
33. Teichrow, J. ; and Horstkotte, E. : Fuzzy-C compiler User's manual, Togai InfraLogic Inc., April 1989.

N92-22695

NEURAL NETWORK TECHNOLOGIES

James A. Villarreal
Software Technology Branch/PT4
NASA Johnson Space Center
Houston, TX 77058

INTRODUCTION

A whole new arena of computer technologies is now beginning to form. Still in its infancy, neural network technology is a biologically inspired methodology which draws on nature's own cognitive processes.

Many problems of current interest (e.g., failure detection and isolation, machine vision and robotics) require the extraction of useful information from a high-volume stream of data from many different sources. The problem is compounded by the fact that often the relation between the input data and the desired output is exceedingly complex, or even unknown. It is therefore necessary to develop advanced modeling tools which are capable of "learning" the dynamics of a problem by exposure to examples of typical system behavior. In many situations, neural networks have proven to be an appropriate modeling tool. Successful applications of neural network technology include modeling and data fusion problems, classification of visual and acoustical patterns, robotics, optimal structural design, recognition of hand written characters, speech synthesis, prediction of solar activity, financial forecasting, modeling of fuel consumption for nuclear reactors, spectrographic analysis, and many others.

The Software Technology Branch has provided a software tool, NETS (Neural Execution and Training System), to industry, government, and academia to facilitate and expedite the use of this technology. NETS is written in the C programming language and can be executed on a variety of machines. Once a network has been debugged, NETS can produce portable C source code which implements the network. This code can then be incorporated into other software systems. This paper will describe various projects currently under development with NETS and the anticipated future enhancements to NETS and the technology.

NEURAL NETWORK SIMULATION TOOLS

Network Execution and Training Simulator (NETS)

NETS (Network Execution and Training System) is a portable C-based simulation of the popular back-propagation neural network which can run on most computers. NETS is intended as a tool for anyone interested in exploring the use of neural networks. It is easy to use, but sufficiently rich in features that it has been used extensively in support of our in-house research. NETS is also available through COSMIC. NETS is currently in use at 288 government sites.

Neural Network Environment on a Transputer System (NNETS)

Due to extreme computational demands which arise in training of ANS, a neural network station was designed and built to be hosted by a network of Transputers. This work station software is currently in the process of being released to Cosmic Computing. The work station was designed to allow for the simulation of a wide range of neural network architectures. There is a sophisticated graphics interface to allow the user to visualize the progress of the network toward convergence. The network of 40 Transputers yields the performance of a super-computer at a fraction of the cost.

NEW TOOLS FOR ANS RESEARCH

Space-Time Neural Networks (STNN)

A spin-off of a project to monitor space shuttle main engine performance with neural networks resulted in the development of an innovative neural network architecture especially useful at sequence discovery and nonlinear filtering. This neural network borrows concepts from associative memory and filtering techniques. In its present form, the STNN dynamics adapt and learn sequences of finite lengths. Currently, efforts are underway to extend the STNN to sequences of arbitrary lengths. This technique is presently undergoing patent issuance.

FASTER TRAINING ALGORITHMS

A major barrier to the use of artificial neural systems is the potentially large amount of processing power required to accomplish the training process. Therefore much research is directed toward the improvement of training methods. This research has resulted in some breakthrough results which are now described.

The Difference Optimized Training Scheme for the Neocognitron Neural Network (DOTS)

The neocognitron is an artificial neural model which was intended to simulate some of the functions of the optic nerve and visual cortex. As originally conceived, the network would "learn" to recognize objects by repeated exposure, a process which often required hundreds of passes to achieve satisfactory results. The DOTS algorithm partitions the training images into a large number of possible templates, and intelligently selects a small collection of templates which describe the training data. From these templates, the synaptic connection strengths may be easily constructed. The time requirements for DOTS are much less than those for the iterative algorithm.

Accelerated Training Method for the Back-Propagation Neural Network

This back-propagation network has become an extremely popular tool for modeling complex functions and pattern recognition. For large networks, the standard training algorithm may be extremely time-consuming if for no other, because the network may have thousands of weights to be determined. The accelerated training algorithm exploits the fact that the input weights for the network (often the vast majority) can be represented as a linear combination of the input vectors. This representation of the input weights can result in a big reduction in the number of optimization variables which must be considered, and thus in a significant savings of time. The degree of effectiveness of the algorithm is strongly dependent on the particular problem under consideration.

Pre-Processing Tool

Often the data which would be natural to present to a neural network is highly redundant. In such a case, an orthogonal basis is extracted from the vectors which comprise the input space. Vectors having the greatest information content are extracted first, and it is possible to determine when there is no benefit in extracting more vectors. The input vectors are projected along the orthogonal basis vectors and the projection coefficients are given to the network instead of the original input. For the problem which suggested the development of this tool, it was able to replace a network with over 15,000 weights with a network having 19 weights. The savings in training and execution times is apparent.

APPLICATIONS

Our anticipated applications for neural networks will be as pattern classifiers, content addressable memories and general non-linear adaptive models. The use of neural networks as part of computer vision systems for inspection, retrieval and rescue is contemplated. Prototype networks have been built, trained and

tested on several hard problems including finding concentration from reading IR spectrograms, prediction of solar activity, speech synthesis from text, and computer vision applications.

Posture Maintenance Control

A joint STB and Life Sciences Directorate project is currently underway to investigate the uses of neural network technology to discover specific sites in the central nervous system which are key for the maintenance of posture. The network will be used to analyze posture platform data collected from patients with lesions in the central nervous system and post and pre flight data from astronauts. It is hoped that neural networks will be able to formulate the appropriate correlations in the data. The STB recently acquired a commercial neural network simulator from SAIC to support this project.

Control of postural equilibrium is a basic physiological function that is mediated by multiple neurosensory feedback systems. Comparisons of pre- and post-flight data have shown alterations thought to be caused by sensory adaptation to microgravity. These alterations will affect the ability of astronauts to perform certain tasks (such as emergency egress immediately postflight). The goal of this project is to develop a neural network model of postural equilibrium control that can map sensory input signals from the visual, visual, and proprioceptive systems into appropriate motor control strategies for maintenance of stable, upright posture.

Sunspot Prediction

Since 1834 reliable sunspot data has been collected by the National Oceanic and Atmospheric Administration (NOAA) and the U.S. Naval Observatory. Recently, considerable effort has been placed upon the study of the effects of sunspots on the ecosystem and the space environment. The ability to predict sunspot activity plays an increasingly important role in both earth and space endeavors. This effort has demonstrated the effectiveness of a neural network (4.5%) RMS error) to model and predict sunspot activity.

Tether Control, Skip Rope Identification, and Remediation

Tethers offer many interesting and useful applications for space-borne systems: electrical power generation, substantial fuel savings, upper atmosphere research, reboost mechanisms, etc. However, the control laws for such systems are non-linear and complex. This task is investigating the replacement of the conventional control system with a fuzzy logic controller. Thereby allowing utilization of imprecise or inexact sensors currently not possible with conventional control systems.

A space shuttle flight scheduled for 1992 will attempt to prove the feasibility of operating tethered payloads in earth orbit. Due to the interaction between the Earth's magnetic field and current pulsing through the tether, the tethered system may exhibit a circular transverse oscillation referred to as the "skiprope" phenomenon. Effective damping of skip rope motion depends on rapid and accurate detection of skiprope magnitude and phase. Because of non-linear dynamic coupling, the satellite attitude behavior has characteristic oscillations during the skiprope motion. Since the satellite attitude motion has many other perturbations, the relationship between the skiprope parameters and attitude time history is very involved and non-linear. We propose a Space-Time Neural Network (STNN) implementation for filtering satellite rate gyro data to rapidly detect and predict skiprope magnitude and phase. Training and testing of the skiprope detection system will performed using a validated Orbital Operations Simulator (OOS) and Space-Time Neural Network software developed in the Software Technology Branch (STB) at NASA's Lyndon B. Johnson Space Center.

FUNDING

Other than two Director's Discretionary Funding projects, funding for neural network technology has been difficult to acquire. However, STB has been successful with the use of the SBIR program to fund two neural network related projects:

1) Netrologic (San Diego, CA) investigated the use of neural network technology for the monitoring of the space shuttle main engine (SSME) for failures. This is a full-scale example of a data fusion problem in that data from many different kinds of sensors must be reduced to obtain an estimate of the system's health. During launch, several SSME sensor readings are monitored by Booster Officers in the MCC. Applying data fusion to this system would, in a sense, "fuse" all sensor readings into a smaller, manageable number and allow the utilization of all the sensor readings in place of the 5 currently monitored. Classification into nominal and faulty engine behaviors would be determined from actual SSME test and launch data.

2) Martingale Research Corporation (Allen, TX) is developing a new neural network architecture called the Parametric Avalanche which is designed to implement a solution to the general stochastic filtering problem for non-linear systems in arbitrary noise. This neural network should see applications in the: 1) Tracking and control of nonlinear dynamical systems, such as spacecraft docking maneuvers, 2) Multi-sensor integration for monitoring of complex power systems, and 3) the control of large articulated space structures.

3) Adaptive Processing Concepts - a NASA headquarters project sponsored by Code R is anticipated to begin in FY 93. This major basic research program will be shared between JSC, AMES, and JPL as lead center. The project will focus on neural network theory and architecture development, breakthrough technologies, and advances in neural network hardware to be directed toward 4 applications: non-linear electro-mechanical control, fault diagnosis and remediation, science data reduction, and geophysical data analysis.

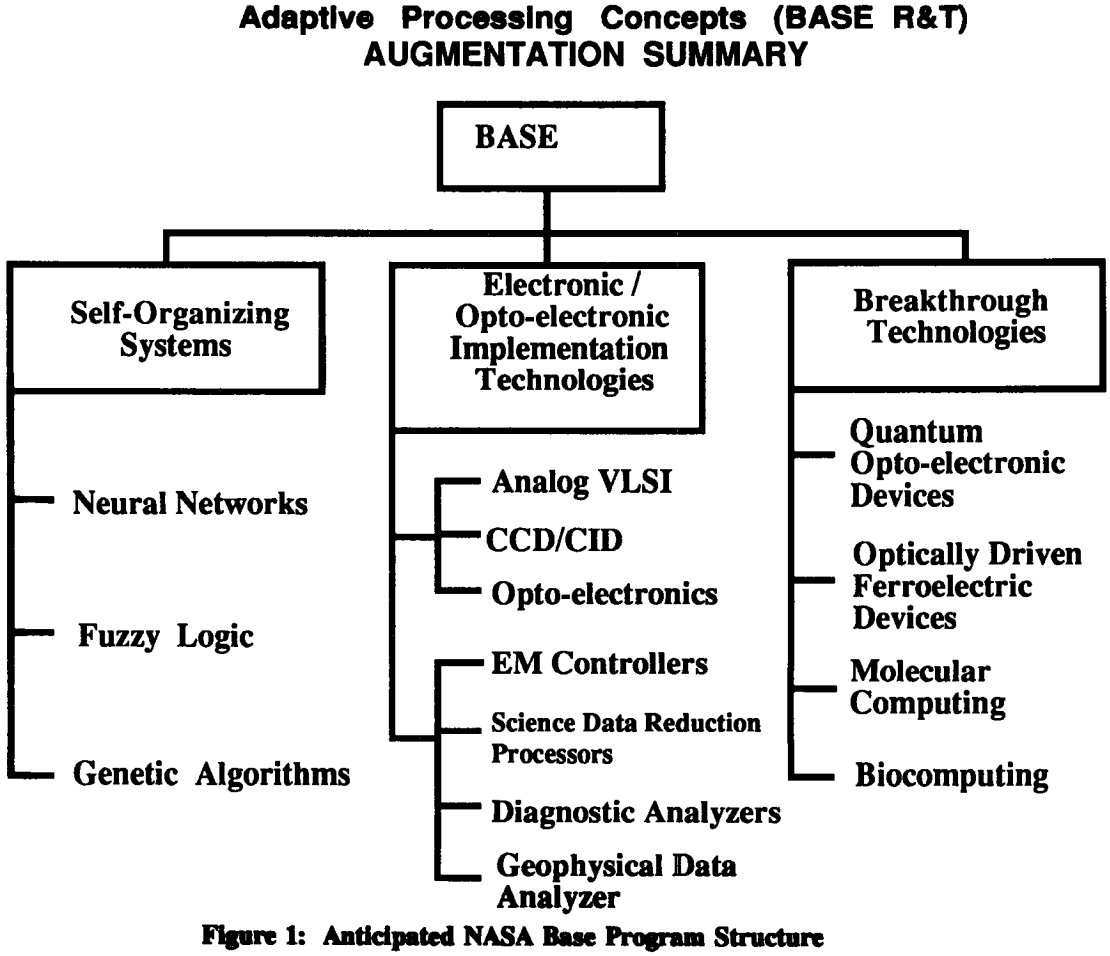


Figure 1: Anticipated NASA Base Program Structure

Measuring Effectiveness and Reliability

Our method for evaluating the performance of a neural network model is inherited from traditional validation procedures used for other modeling techniques (i.e., least squares, polynomial fitting, trigonometric series expansions). Where sufficient data exists, it is split into a training set and a test set. The training set is used to adapt the coefficients in the network. Performance is measured on the test set. This method is extremely reliable since it tests the ability of the network to generalize based on the performance on the test set. The problem lies in the expense and difficulty in producing data for the systems of greatest interest.

Flight Qualification of Artificial Neural Systems

A neural network could be flight qualified only in a situation which would allow the action of the network to be evaluated over a very large set of possible inputs. In such a case the network would have the same degree of reliability as many other types of models which are routinely used flight software. Neural network technology should be verified as any other type of model I.e. through extensive and continued use in simulations until a satisfactory degree of confidence has been achieved.

REFERENCES

1. Copeland, C., Lea, R., Jani, J., and Villarreal, J. "Fuzzy Logic Based Tether Control," *North American Fuzzy Information Processing Society - '91*, Columbia, Missouri, 1991.

2. McIntire, Gary and Villarreal, James, "Design of a Neural Network Simulator on a Transputer Array," *First Annual Workshop on Space Operations Automation and Robotics (SOAR 87)*, NASA CP-2491, pp 111-119, Houston, TX, 1987.

3. Savely, Robert and Villarreal, James, "The Implementation of Neural Network Technology," *IEEE First International Conference on Neural Networks*, pp IV477-IV485, San Diego, CA, 1987.

4. Shelton, Robert O., "A Difference Optimized Training Scheme for the Neocognitron Neural Network," Proceedings of Instrument Society of America Conference, October 1988.

5. Villarreal, J. A. and Baffes, P., "Sunspot Prediction Using Neural Networks," with Paul Baffes, *SOAR '89 - Third Annual Workshop on Automation and Robotics*, Houston, TX, 1987.

6. Villarreal, J. A. and Shelton, R. O., "A Space-Time Neural Network," Second International Conference on Neural Networks and Fuzzy Logic - 1990, Houston, TX, 1990.

N92-22696

FROM BIOLOGICAL NEURAL NETWORKS TO THINKING MACHINES: TRANSITIONING BIOLOGICAL ORGANIZATIONAL PRINCIPLES TO COMPUTER TECHNOLOGY

Muriel D. Ross
Director, Biocomputation Center
NASA Ames Research Center
Moffett Field, CA 94035

ABSTRACT

The three-dimensional organization of the vestibular macula is under study by computer assisted reconstruction and simulation methods as a model for more complex neural systems. One goal of this research is to transition knowledge of biological neural network architecture and functioning to computer technology, to contribute to the development of "thinking" computers. Maculas are organized as weighted neural networks for parallel distributed processing of information. The network is characterized by non-modularity of its terminal/receptive fields. Wiring appears to develop through constrained randomness. A further property is the presence of two main circuits, highly channeled and distributed modifying, that are interconnected through feedforward-feedback collaterals and a biasing subcircuit. Computer simulations demonstrate that differences in geometry of the feedback (afferent) collaterals affects the timing and the magnitude of voltage changes delivered to the spike initiation zone. Feedforward (efferent) collaterals act as voltage followers and likely inhibit neurons of the distributed modifying circuit. These results illustrate the importance of feedforward-feedback loops, of timing, and of inhibition in refining neural network output. They also suggest that it is the distributed modifying network that is most involved in adaptation, memory and learning. Tests of macular adaptation, through hyper- and microgravitational studies, support this hypothesis since synapses in the distributed modifying circuit but not the channeled circuit are altered. Transitioning knowledge of biological systems to computer technology, however, remains problematical.

A major objective of neuroscience research is to understand the relationship between neural geometry and the coding of information. Neural geometry is defined here as the organization from the subcellular to the network and parallel systems level that permits the transfer of information from one site to another, so that some activity, whether that be an abstract thought or a motor response, results. While the goal of this kind of research is, more often than not, to gain knowledge of neural functioning as a basis for understanding normal and diseased states, the potential for spin-offs in computer technology is enormous. For what can emerge from work based in neurobiology is nothing less than a computer with "sense" as well as "knowledge"; that is, a computer that can "learn" and "think".

At the Biocomputation Center at NASA-Ames Research Center, my research team focuses on gaining an understanding of the fundamental properties of a simple mammalian neural network, the vestibular macula of the rat inner ear, as a model for more advanced neural systems. Maculas are linear bioaccelerometers that are organized similarly to engineered accelerometers: they have a test mass that is not tightly attached to the underlying detecting unit while the sensing part is anchored. It is thought that differences in the relative motions of the test mass and the detector during acceleration of the head provide the stimulus to the receptor cells, called hair cells. The signals are transduced into electrochemical activity by the hair cells, and macular neurons inform the central nervous system about the acceleration of the head in three-dimensional (3-D) space. The information, continuously transmitted, is integrated centrally with input from angular bioaccelerometers and other sensory systems, such as the visual and proprioceptive (joint and muscle sense), to keep the organism stable and in correct posture for the task at hand, whether the animal is at rest or in motion.

The sensing part of the linear bioaccelerometer is a neural network that, like retina and other parts of the brain, is organized for weighted, parallel distributed processing of information [1,2]. The test mass (otoconia) is not uniformly distributed over the sensory array, so that incoming signals are not passed identically to all parts of the network. Thus, in contrast to engineered systems, maculas are organized for parallel processing at both the input (test mass) and sensing (neural) levels.

The receptor hair cells of the macular neural network are of two types, I and II (Figure 1). Type I cells are almost completely enclosed by expanded terminals (calyces) of the nerve fibers and communicate only with calyces. Type II cells, in contrast, lie between calyces and synapse directly with them or with their collaterals. Type II cells also are under the influence of calyceal and nerve fiber activity from vesiculated collaterals of calyces and nerve fibers. Additionally, a system of very fine, unmyelinated nerve fibers provides presynaptic endings (efferents, [3]) for calyces, nerve fibers and type II hair cells. This efferent system may perform a biasing function in the network (see below).

One of the major features of the mammalian macula is the non-modular organization of its neural elements. No two terminal/receptive fields are identical. There is a spectrum of terminal field patterns ranging from those with a single calyx springing from the first heminode of a myelinated nerve fiber (M-type) to those with a branched pattern of two or three calyces and a long, unmyelinated segment distal to the heminode (U-type) [4,5]. An intermediate variety has one or two calyces and a short, unmyelinated preterminal segment (M/U-type). Similarly, the receptive fields differ in the number and arrangement of receptor cells. The morphological range from M-type to U-type innervation patterns is reflected in a physiological spectrum of discharge properties from highly irregular to regularly firing [6,7] that encode phasic to tonic head acceleration and positioning.

Our research with Monte Carlo simulations, using a matrix of measurements from the biological data and a random number generator, show that non-identical terminal/receptive fields could develop through a process of constrained randomness [8]. We were able to reproduce the patterns actually observed in the macular neural network by this computer method. Constrained randomness in wiring is very likely one way in which biological neural networks maintain robustness and the ability to degrade gracefully during aging while simultaneously ensuring variability and capability to meet unknown environmental challenges. Still, to my knowledge, constrained randomness has not been included in artificial neural networks and its advantages or disadvantages for computer technology are unexplored.

A further property that is coming under increased scrutiny for potential computer implementation is the fundamental organization of maculas and other biological neural networks into two main circuits. The first circuit consists of an input with relatively direct access to the output neurons while the second circuit distributes information to other cells and modifies the final output of the network. I have called these two circuits "highly channeled" and "distributed modifying" respectively (Figure 1) [8]. However, the two circuits correspond to those described for olfactory system and retina by Shepard [9] as "vertical" and "horizontal", and for cortex by Schmitt [10] as "direct" and "local".

The distributed modifying circuit involves cells that conform to the idea of interneurons. Interneurons are local cells that are inserted into conducting pathways to modify output of the transmission lines, but that do not conduct information over long distances. A distributed modifying circuit may include a single interneuron, as it does at the first stage of retinal processing, where horizontal cells take on this function; or it may consist of several interneurons and more complicated circuitry as it does in cerebral cortex. Although it is a novel use of the term "interneuron", type II hair cells should be considered to belong to this class since they are inserted into feedforward-feedback loops in the distributed modifying circuit. They function as a mixture of receptor cell and interneuron. In the retina, for example, a comparable cell would be a hybrid photoreceptor/horizontal cell.

In the macula, a portion of the distributed modifying circuit may have a biasing effect on neural elements of the network, as mentioned briefly above. This part consists of the highly vesiculated, button-like endings (boutons) that terminate on type II hair cells, calyces, intramacular nerve fiber branches, and collaterals (see Figure 1). Although all nerve fibers of this system are cholinergic [11], the morphology of the synapses made by their terminals suggests that they may have a dual function in the network that depends on subsynaptic features. For example, the boutons terminating on type II hair cells end opposite or near subsynaptic cisterns, which are part of smooth endoplasmic reticulum network of the cell. They thus correspond to c synapses [12] and likely function to hyperpolarize (inhibit) the cell via calcium release from the cisterns and resulting potassium extrusion [13] (see supporting evidence provided by Ohmori [14] for vestibular hair cells). (Efferent-type intramacular collaterals are considered here to hyperpolarize the type II hair cells because they, too, end opposite subsynaptic cisterns.) That some of the c synapses are involved in highly local processing is apparent because

~20% to 30% of them are reciprocal (there are both pre- and post-synaptic structures at the hair cell membrane) [8]. Synapses formed by other terminals on calyces, nerve fiber branches and collaterals are asymmetric synapses. The asymmetric synapses should have a depolarizing (excitatory) effect on these neural elements, keeping them at some state of activation depending on the number of synapses collectively transmitting at any given time. It is most interesting that many of these terminals form more than one junction with the postsynaptic element. Some terminals synapse with both a type II hair cell and a calyx where, on morphological grounds, they should bring about opposite effects. How common biasing circuits are in neural systems is unclear, but a comparable intrinsic cholinergic circuit has been described for optic tectum [15] where the circuit was shown experimentally to enhance retinotectal transmission.

Thus, the type II hair cell is inserted into feedforward-feedback loops in the network and some of its synaptic sites are morphologically organized to support highly localized neural processing. The type II cell also appears to function against both a transient inhibition from feedforward collaterals and a continuous biasing inhibition from the efferent terminals. This is in contrast to the type I cell, which is isolated from external influences except for those that might be transmitted through the calyx.

In order to begin to test the significance of the distributed modifying circuit, of local processing and of biasing in neural computation, we conducted a compartmental modeling study of collaterals using precise measurements from our serial sections and software called NEURON [16]. The question posed for this initial investigation was whether collaterals of differing morphologies would yield significantly different voltage contributions to neural activity, demonstrating a relationship between neural geometry and functioning.

Several collaterals with different morphologies were examined. Each terminal was converted into a matching geometric form which was subdivided into major segments depending on the taper rate (cylinder, frustrum, etc.) of each part. Each segment was partitioned into 30 compartments for modeling purposes, with the number of compartments deemed sufficient for the model determined empirically. The collaterals were all considered to arise from a nerve fiber with a diameter of 2.5 mm and a length of 2 l (1118 mm), divided into 200 compartments. Voltage changes were monitored at the input site (the distal end of the process), the base of the process, at a distance of 33.5 mm and at 145 mm along the nerve fiber branch. Following the initial study, simulations were conducted in which stem diameter was kept constant at 0.2 mm but the length was increased by factors of two, (or by the square root of two for the smaller lengths) from 0.5 mm to 16 mm. In other simulations, the stem length was kept constant at 0.8 mm (this was a real dimension), but the diameter was changed in the same manner from 0.07 mm to 1.6 mm.

Briefly, the results showed that when the collaterals and nerve fibers were considered to be passive and afferent functionally, the voltage changes at the base of the stem are very sensitive to changes in stem diameter but not as sensitive to changes in length. Interestingly, the size or shape of the head did not matter as much in terms of voltage changes delivered to the nerve fiber branch. Moreover, increasing the diameter of the stem not only resulted in increased voltage delivered to the base, but also shortened the time to peak delivery. The conclusion reached was that small changes in diameter of the collateral, anywhere this might occur, alter both the amount of voltage and its timing of delivery to the base of the collateral. This would mean that all afferent collaterals are not equivalent with respect to their influence on a calyx, or on the activity of other collaterals arising from nearby on the calyx. These results, based on biological data, support the concept that geometry affects both the timing of input and the magnitude of the voltage changes at the spike initiation zone in biological neural networks.

Results of similar simulations in which the collaterals were considered to be efferent in type showed that, in all cases, voltage changes were delivered to the type II hair cell quickly and without much diminution. Feedforward collaterals would thus seem capable of influencing type II cell activity, depending on the dynamic state of the calyx or nerve fiber branch, quickly and with minimal loss of current at the source. In the electronic sense, the efferent collateral is a voltage follower.

The degree of inhibition a type II experiences from moment to moment is a dynamic parameter. The level of inhibition depends on the timing of the *total* feedforward input from all the calyces and nerve fiber branches from which the type II cell receives efferent-type terminals and collaterals. Also, any individual calyx *simultaneously* regulates the transient inhibition of *all* the type II hair cells with which it communicates, whether they feed back to that calyx or to another. The dynamics of this inhibition in turn determine the amount of excitatory feedback provided to the calyces, since this depends on how much a type II cell's own excitatory activity overcomes imposed inhibition. Because of the short distances involved, the voltage changes in one part of the terminal array affect the entire array, so that there must be continuous fluctuation in transmembrane voltage, with the output reflecting the differing neural geometries of the terminal/receptive fields. This is an analog system that converts its output into a digital signal (discharge).

Our simulations, then, begin to tell us something about the possible importance of feedforward-feedback loops in biological neural networks, the necessity of timing, and the need for inhibition to control and refine output. The role of reciprocal synapses and of highly localized processing remain to be explored. Assuming that our interpretations of the basic circuitry are correct, it would be predicted that it is the distributed modifying circuit that would show most plasticity and adaptability to an environmental change. According to Shepard [9], it is this circuitry that would also be most involved in memory and learning.

This concept, if correct, has great relevance to the development of advanced, "thinking" computers. Although there is currently a great deal of research into neural plasticity and learning [see 17], the central idea that there are two fundamental circuits with one of them more susceptible than the other to adaptive change has, to my knowledge, not been specifically tested. However, a proposed explanation for neural modulation during learning in *Aplysia* [18] utilizes two basic circuits, one that is direct, or highly channeled, and the other that is routed through a modulating pathway. Additionally, the cerebellar circuitry is a classic example of a system with a highly channeled input (the climbing fiber) and a distributed modifying one (the parallel fibers), and research indicates that the latter circuit is the one that is more adaptive [19].

In contrast to many other neural networks, maculas are a perfect system to serve as a testbed for the concept of dual circuitry with non-equivalent plasticity. The macular network is simple compared to other more highly evolved systems, and it responds to transient linear acceleration and to steady-state gravitational stimuli. The gravitational environment can be manipulated through the use of the space shuttle and ground-based centrifuges. Additionally, changes in macular synaptic or collateral connectivities in response to an increase or decrease in gravitational bias can be determined more readily than ever before using computer-aided counting and reconstruction methods.

Early, preliminary findings in maculas from animals flown on SLS-2 for 9 days and from others subjected to 2-g centrifugation for two weeks indicate that synapses in type II hair cells decrease by ~30% in hypergravity and increase by ~25% as a consequence of spaceflight. Synapses in type I cells were not significantly affected when compared to controls. If these results hold up when the study is completed, we shall have demonstrated that the macula can adapt to an environmental change in linear acceleratory input, bidirectionally. Moreover, the conclusion will be almost inescapable that feedforward collaterals help regulate the excitatory output of the type II cell, including synapse formation and degradation. This would mean that some mechanism exists, even in these simple cells, for gene expression to alter the cell's structure in response to a differing pattern of activation. In other, central neural tissues, this is described as a form of learning (see discussions in [17]).

This brings us to another important issue: whether "memory" and "learning" are primitive attributes of every neuron finding highest expression in the human brain due to the enormous increase in association cortex, a distributed modifying type of neural circuitry. This would mean that although each neuron individually is still essentially an automaton, collective expression of individual mechanistic capablities, genetically assured, underpins emotional response, rational thought and creative ability.

How to interact best with industry to transition these and other insights emerging from biological research to computer technology remains problematical. My experience is that mechanisms to facilitate transitioning from science to industry are either not in place or do not work well. The vision of commercial payoff may be too short-sighted, or our hopes of assisting industry may be premature. What is required is closer interaction between theoreticians, experimentalists and implementers, but this kind of merger thus far appears easier to bring about in an individual laboratory than between science and industry, which often have disparate goals.

Nevertheless, I am certain that computer modeling of three-dimensional neural networks mimicking properties of biological systems will play an essential role in achieving knowledge of how to construct more intelligent computers. Indeed, one exciting prospect of our 3-D modeling effort is that one can begin with a simple system such as the vestibular macula and, by adding interneurons, build up to a retina or more complex cortex, even to parallel circuits that include an associative cortex, to learn what functional advantage is achieved by splitting cellular functions and increasing the complexity of distributed modifying circuits. While these simulations could aid in the understanding of vestibular maculas and of retina, they offer the additional promise of theoretical insights advantageous to computer technology and to the production of more sophisticated "thinking" machines. Rarely has there been so perfect an opportunity to reap biological, biomedical and technological advantage from a single nexus of science and industry; unfortunately for both players, we are still learning how to bring this bonding about.

REFERENCES

1. Ross MD. Anatomic evidence for peripheral neural processing in mammalian graviceptors. Aviat Space Environ Med 1985; 56: 338-43.

2 Ross MD. Morphological evidence for parallel processing of information in rat macula. Acta Otolaryngol (Stockh) 1988; 106: 3-9.

3. Engstrom H. On the double innervation of the sensory epithelia of the inner ear. Acta Otolaryngol (Stockh); 49: 109-18.

4. Ross MD, Rogers CM, Donovan KM. Innervation patterns in rat saccular macula. Acta Otolaryngol (Stockh) 1986; 102: 75-86.

5. Ross MD, Cutler L, Meyer G, Lam T, Vaziri P. 3-D components of a biological neural network visualized in computer generated imagery. I. Macular receptive field organization. Acta Otolaryngol (Stockh) 1990; 109: 83-92.

6. Fernandez C, Goldberg JM. Physiology of peripheral neurons innervating otolith organs of the squirrel monkey. I. Response to static tilts and to long-duration centrifugal force. J Neurophysiol 1976; 39: 970-84.

7. Tomko DL, Peterka RJ, Schor RH. Responses to head tilt in cat eighth nerve afferents. Exp Brain Res 1981; 41: 216-221.

8. Ross MD, Cutler L, Doshay D, Cheng R, Naddaf A. A new theory of macular organization based on computer-assisted 3-D reconstruction, Monte Carlo simulation and symbolic modeling of vestibular maculas. Acta Otolarygol (Stockh) 1991; Suppl. 481: 11-14.

9. Shepard GM. The olfactory bulb as a simple cortical system: Experimental analysis and functional implications. In: Schmitt FO, ed. The Neurosciences: Second Study Program. New York: Rockefeller Univ Press, 1970: 539-52.

10. Schmitt FO. The role of structural, electrical and chemical circuitry in brain function. In: Schmitt FO, Worden FG, eds. The Neurosciences: Fourth Study Program. Cambridge, MA: MIT Press, 1979: 5-20.

11. Iurato S, Luciano L, Pannese E, Reale E. Histochemical localization of acetylcholinesterase (AchE) activity in the inner ear. Acta Otolaryngol (Stockh) 1971; Suppl 279.

12. Conradi S. Ultrastructure and distribution of neuronal and glial elements on the motorneuron surface in the lumbosacral spinal cord of the adult cat. Acta Physiol Scand 1969; 332: 85-111.

13. Connaughton M, Priestly JV, Sofroniew MV, Eckenstein F, Cuello AC. Electron microscopic immunocytochemistry for choline acetyltransferase, substance P and enkephalins using monoclonal antibodies. Neuroscience 1986; 17: 205-24

14. Ohmori H. Mechanoelectrical transduction in the chicken hair cell. Acta Otolarygol (Stockh) 1991; Suppl 481: 1-4.

15. King WM, Schmidt JT. A cholinergic circuit intrinsic to optic tectum modulates retinotectal transmission via presynaptic nicotinic receptors. In: Wolpaw JR, Schmidt JT, Vaughan TM, eds. Activity-Driven CNS Changes in Learning and Development. New York, NY: Ann NY Acad Sci, 1991: 363-367.

16. Hines M. A program for simulation of nerve equations with branching geometries. Int J Bio-med Comput 1989; 24: 55-68.

17. Wolpaw JR, Schmidt JT, Vaughan TM, eds. Activity-Driven CNS Changes in Learning and Development. New York, NY: Ann NY Acad Sci, 1991, 399 pp.

18. Byrne JH, Baxter DA, Buonomano DV, et al. Neural and molecular bases of nonassociative and associative learning in *Aplysia*. In: Wolpaw JR, Schmidt JT, Vaughan TM, eds. Activity-Driven CNS Changes in Learning and Development. New York, NY: Ann NY Acad Sci, 1991: 124-149.

19. Greenough WT, Anderson BJ. Cerebellar synaptic plasticity: Relation to Learning versus neural activity. In: Wolpaw JR, Schmidt JT, Vaughan TM, eds. Activity-Driven CNS Changes in Learning and Development. New York, NY: Ann NY Acad Sci, 1991: 231-247.

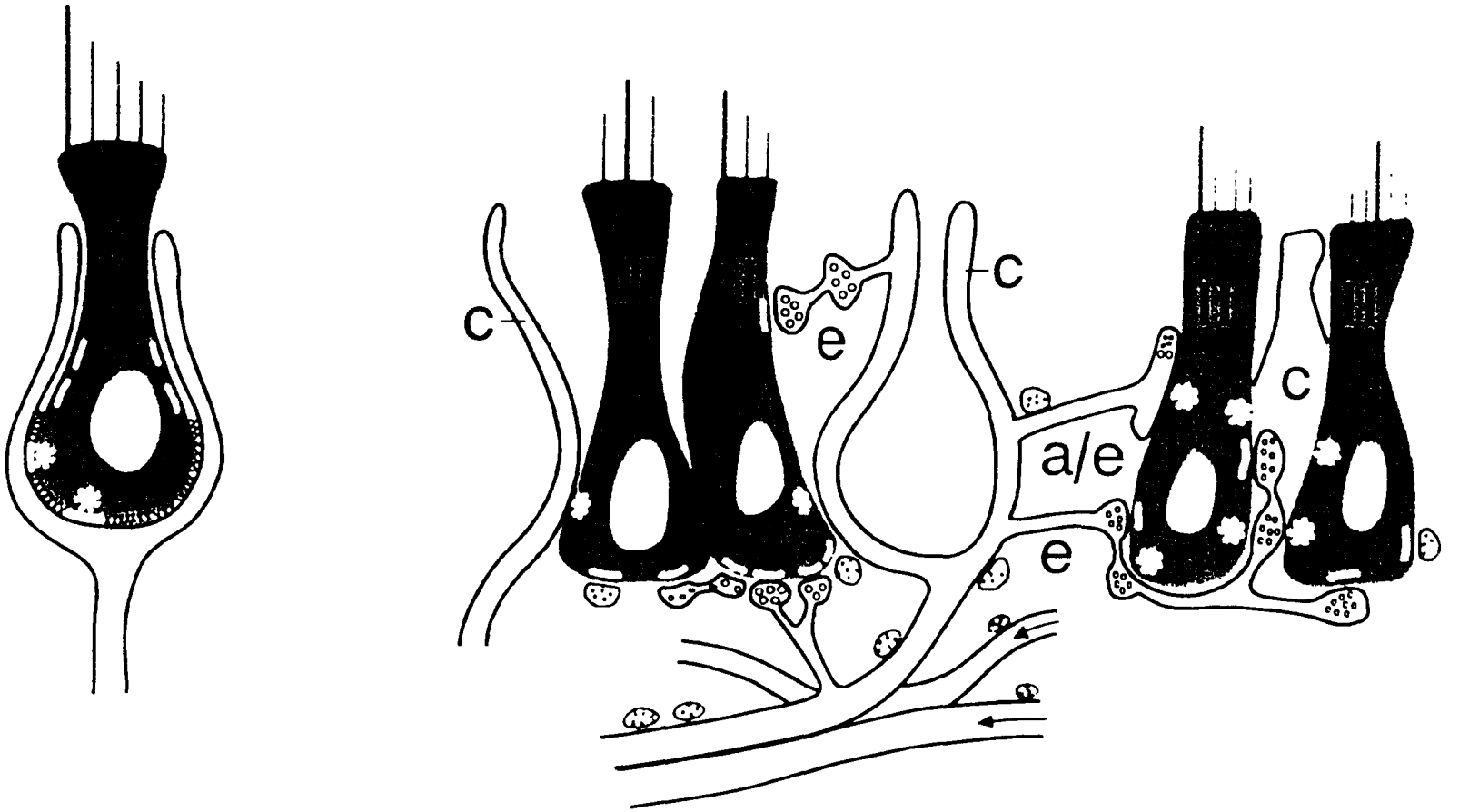


Figure 1. This diagram illustrates the two circuits, highly channeled (left) and distributed modifying (right) present in the vestibular macula. In the channeled circuit, input coming to the type I cell (I) is transduced and transmitted only to the calyx (C) nerve ending. All other parts of the macular neural network belong to the distributed modifying (modifier) circuit. Input to type II cells (II), once transduced, is subject to modification from calyceal and nerve fiber feedforward collaterals (e, efferent and a/e, mixtures of afferent and efferent morphologies) and to biasing from other vesiculated boutons (button-like endings, unlabeled). Type II cells distribute their output to more than one calyceal terminal by directly synapsing with a calyx (asterisk, far left, distributed modifying circuit) or by afferent collaterals (only an a/e collateral is drawn here). This feedback modifies the output that is sent from the macula to the central nervous system Arrows (lower center) indicate the direction of output flow. In the drawing, asterisk-like structures indicate afferent synapses and rod-like structures represent subsynaptic cisterns (see text).

BIOTECHNOLOGY

(Session D3/Room C1)

Thursday December 5, 1991

- **The Microassay on a Card -- A Rugged, Portable Immunoassay**
- **Detection of Small Molecules with a Flow Immunosensor**
- **Nucleic Acid Probes in Diagnostic Medicine**
- **The Rotating Spectrometer: New Biotechnology for Cell Separations**

N92-22697

THE MICROASSAY ON A CARD - A RUGGED, PORTABLE IMMUNOASSAY

David Kidwell, PhD
Code 6177
Naval Research Laboratory
Washington, DC 20375

ABSTRACT

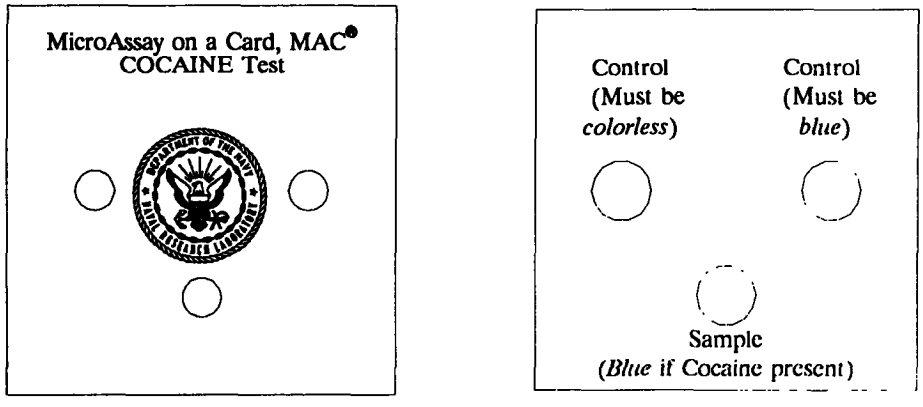
The Microassay on a Card (MAC) is a portable, hand-held, non-instrumental immunoassay that can test for the presence of a wide variety of substances in the environment. The MAC is a simple device to use. A drop of test solution is placed on one side of the card and within five minutes a color is developed on the other side in proportion to the amount of substance in the test solution with sensitivity approaching 10 ng/ml. The MAC is self-contained and self-timed, no reagents or timing is necessary. The MAC may be configured with multiple wells to provide simultaneous testing for multiple species. As envisioned, the MAC will be employed first as an on-site screen for drugs of abuse in urine or saliva. If the MAC can be used as a screen of saliva for drugs of abuse, it could be applied to driving while intoxicated, use of drugs on the job, or testing of the identity of seized materials. With appropriate modifications the MAC also could be used to test for environmental toxins or pollutants.

INTRODUCTION

The Naval Research Laboratory has developed a rapid, field-portable assay for substances in the environment that provides a positive signal in the presence of the substance. The assay relies upon antibody-antigen recognition of the compound of interest, amplification of this recognition, and subsequent visual indication of the presence of the compound. For use, 50 µl of the test fluid is placed in three wells on one side of the MAC. The test fluid is drawn into the MAC by capillary action. No additional reagents are added to the test fluid and all timing is controlled by diffusion inherent in the design of the MAC. The front and back views of the MAC slide are shown in Figure 1.

Depending upon the application, *i.e.* testing of solid or liquid samples, the MAC may be packaged in one of two ways. For testing of solid samples, the product consists of three components: a sealed vial of aqueous buffer, a transfer bulb, and the MAC slide. For testing of liquid samples, only the MAC slide is used.

Figure 1 - Front and Back Views of the MAC (Actual Size is 2" Square)



 PRECEDING PAGE BLANK NOT FILMED

The following steps are used to test solid samples:

1. The vial of aqueous buffer solution is broken open and a small amount (< 1 mg) of the solid sample is added using the transfer bulb.
2. One drop of the solution is placed in each of the three wells of the MAC, also using the transfer bulb. The fluid is transferred through the MAC via capillary action with the flow rate determined by the semipermeable membrane incorporated in the device.
3. After two to five minutes (the timing is not critical), the MAC may be turned over and the results observed with the unaided eye.

In the current design, three spots will be present. Two spots are controls, one of which is a positive control, that must be blue and the other is a negative control, that must be colorless. The presence of a blue spot in the test well indicates the presence of the test substance. No reagents must be added to the test fluid; all timing is controlled by the design of the MAC.

PRINCIPLE OF THE ASSAY

Basic Design of the MAC

The MAC consists of a multi-layer slide that may be configured in one of two general modes: displacement and competitive. Figure 2 shows a schematic side view of the MAC. From left to right, the first part is a hydrophobic layer containing the wells which would be produced out of a suitable plastic material, such as polyethylene. The test fluid is placed in the resultant wells and is held by surface tension. Depending upon the mode of operation, the MAC may have either one or two inserts inside the well. The number of inserts and their function are given in Table 1. Two of the major modes of the design of the MAC are discussed below.

Figure 2 - Schematic Side View of the MAC

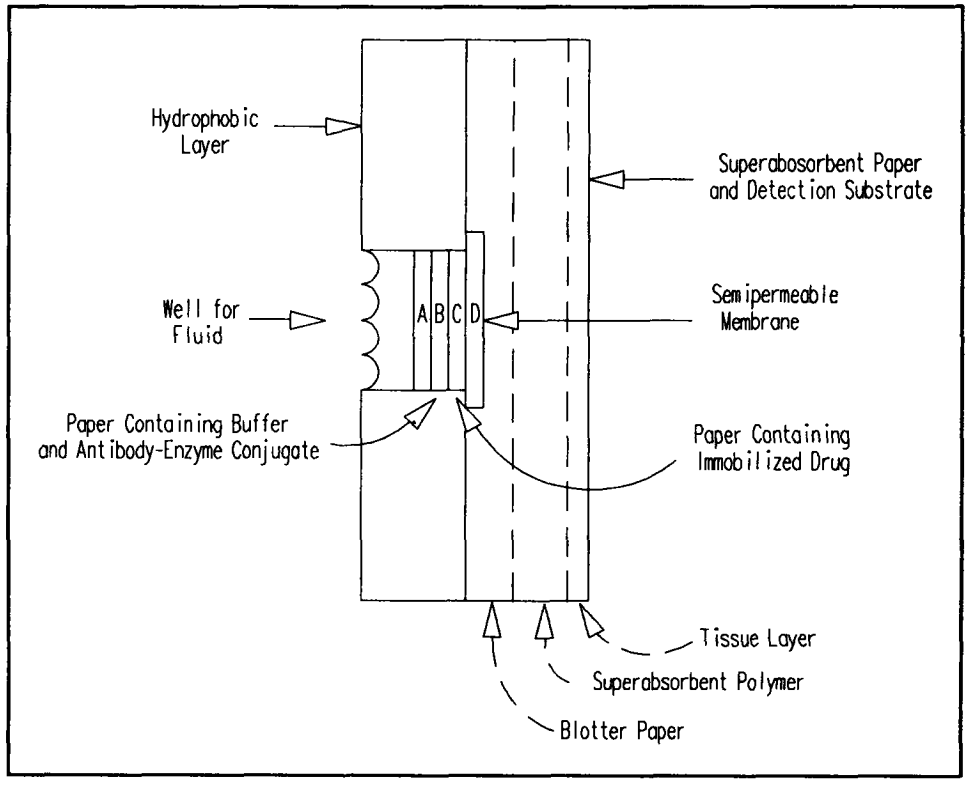


Table 1 - Arrangement of Reactive Species in Well

Type of Assay	Layer A	Layer B	Layer C	Layer D
Displacement 1.	optional	-	antibody-enzyme conjugate + immobilized substance	membrane
2.	optional	-	-	antibody-enzyme conjugate + substance bound to membrane
3.	optional	-	substance-enzyme conjugate + immobilized antibody	membrane
4.	optional	-	-	substance-enzyme conjugate + antibody bound to membrane
Competitive 5.	optional	antibody-enzyme conjugate	substance bound to surface	membrane
6.	optional	substance-enzyme conjugate	antibody bound to surface	membrane
7.	optional	antibody-enzyme conjugate	-	substance bound to membrane
8.	optional	substance-enzyme conjugate	-	antibody bound to membrane

In all modes of operation, there exists the semipermeable membrane layer and the detection layers. The semipermeable membrane controls the timing of the fluid flow by the number of pores and their size. The timing is set such that the test fluid is drawn through the buffer and antibody layers in approximately two minutes. Other timing may be implemented as shown in Table 2.

Table 2 - Timing for Polycarbonate Membranes

Pore Size (μm)	Time (secs)
0.015	∞ (does not draw through)
0.1	800
0.2	135
0.4	68

The detection layer consists of three sub-layers. One is blotter paper that serves as a backing that provides a white background for the developed colored spot. The other sub-layer is a super-absorbent polymer matrix which is impregnated with the substrate for the enzyme. The super-absorbent polymer permits larger liquid samples to be applied and therefore increased sensitivity. The last sub-layer is a tissue paper layer that protects the super-absorbent polymer against abrasion and becomes transparent when wetted by the test fluid.

The advantages of superabsorbent polymers for use in the MAC are many and a demonstration of their use as enzyme detection media has been published.[1] Superabsorbent polymers consist of either salts of polyacrylic acid or grafted acrylic acid on a starch backbone. These polymers can absorb up to 2000 times their weight in water.[2] The advantages of superabsorbent polymers as a reaction medium are: 1) These polymers become hydrated in under three seconds. 2) Once hydrated, the polymer forms a colorless and clear gel. This allows any color forming reaction to be observed throughout the gel which increases the sensitivity. 3) Little control of the amount of fluid is needed. Excess fluid will diffuse out of the reaction area and no color will be produced. 4) Once formed, the gel is restrictive to fluid and molecular diffusion. Therefore, only the enzyme and substrate in the vicinity of the enzyme will react; diffusion of material from other areas is prevented. Also, diffusion of reactants from the detection area is limited. Thus immobilization of the detection reagents is not necessary. 5) The polymers are weak acids and self-buffering around pH 7.

DESIGN OF THE INSERTS IN THE WELL

The number of inserts in the test well and their configuration determines the mode of operation of the MAC; all other components are identical. There are two main modes of operation: Displacement and competitive.

Displacement Mode of Operation of the MAC

In the displacement mode, only one insert is present in the well (optionally two may be present with one containing the buffer). An antigen is chemically attached to the insert and after appropriate manipulation, an enzyme-conjugated antibody is pre-absorbed to the antigen (see Figure 3). When the fluid passing though the slide contains the antigen that the antibody recognizes, some of the enzyme-conjugated antibodies will be displaced from the immobilized antigen. The fluid will carry the displaced enzyme-conjugated antibody into the substrate layer where the enzyme will act on the substrate producing a color. If no antigen is present in the test fluid, no enzyme-antibody conjugate will be displaced and therefore no color will be produced in the substrate layer. The color is proportional to the amount of enzyme displaced and, therefore, proportional to the antigen concentration in the test fluid. (A darker color means more drug is present.)

Competitive Mode of Operation of the MAC

In the competitive mode, two inserts are present in the well. On the first insert is adsorbed enzyme-conjugated antibody and chemically attached antigen is contained on the second insert (see Figure 4). When fluid is passed though the slide, the fluid dissolves the enzyme-conjugated antibody and carries it through the MAC. If no antigen is present in the test fluid, all the enzyme-conjugated antibody would bind to and be trapped in the second layer. Thus no enzyme would appear in the detection layer and no color would develop. Conversely, if antigen was present in the fluid, all the binding sites on the antibody would be blocked and little or no enzyme-conjugated antibody would be trapped in the second layer. Thus a strong color would develop in the detection layer. In this manner, the color in the detection layer is proportional to the amount of enzyme present and, therefore, proportional to the antigen concentration in the test fluid. (A darker color means more drug is present.)

Figure 3 -Configuration of the MAC for a Displacement Immunoassay

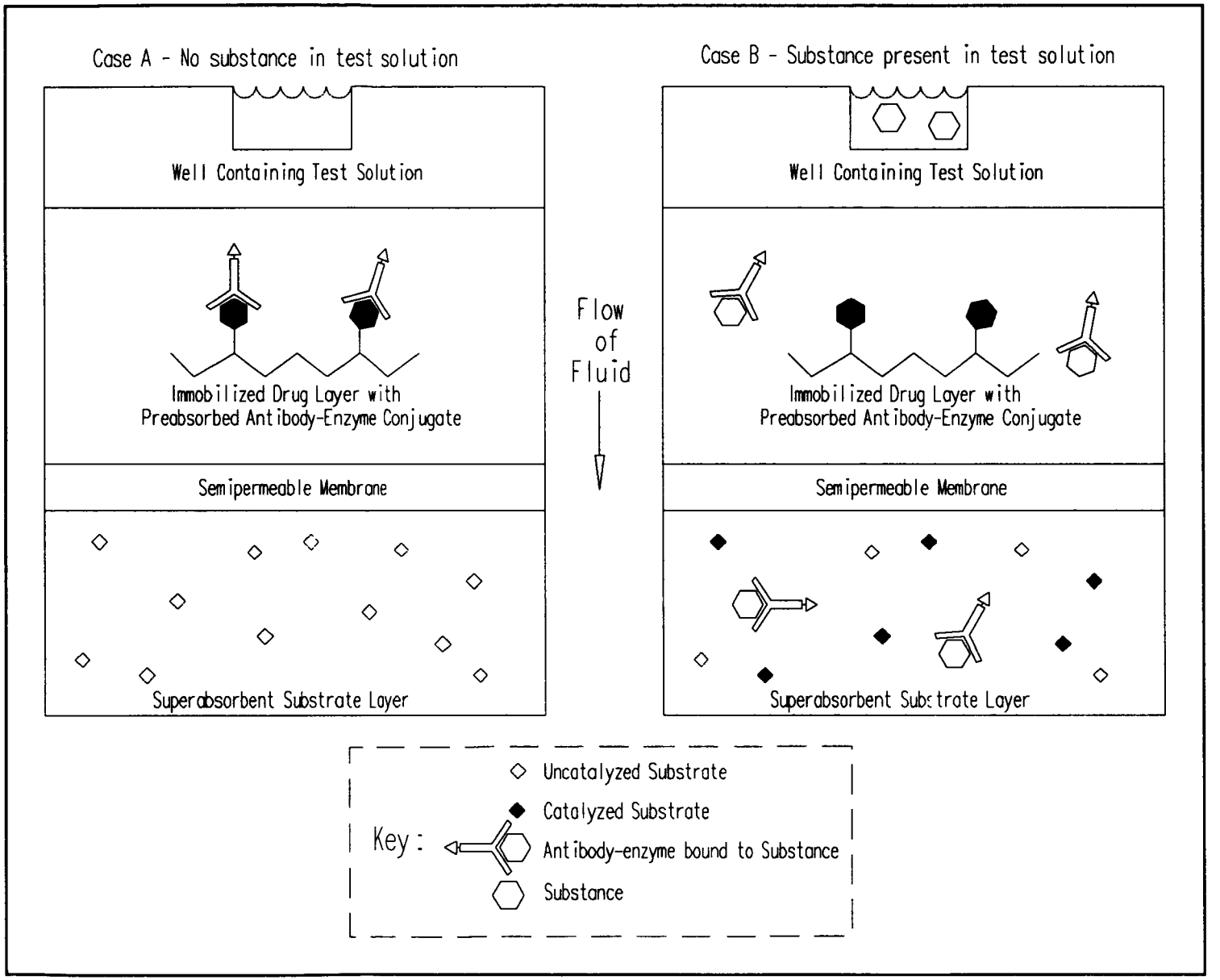
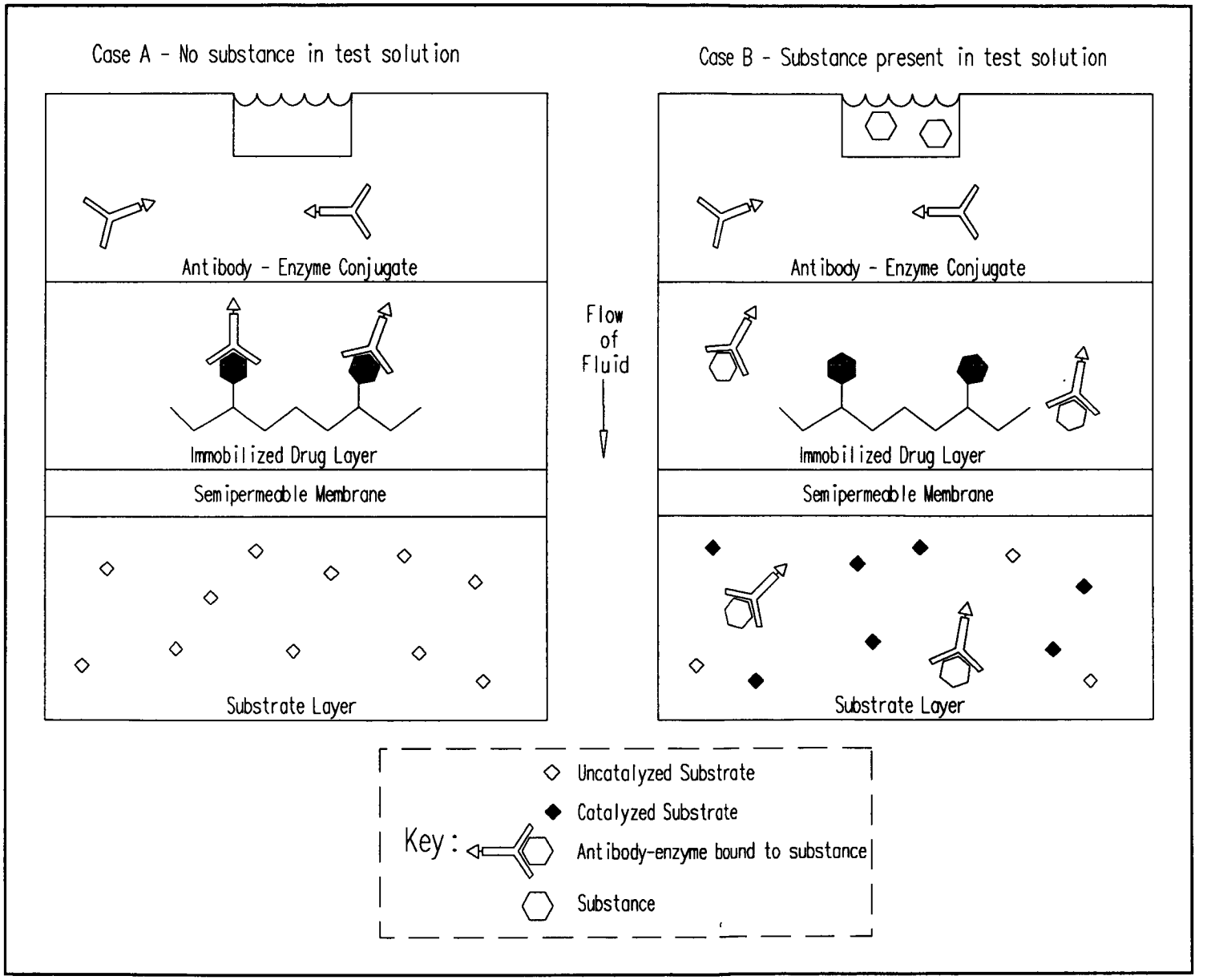


Figure 4 - Configuration of the MAC for a Competition Immunoassay



Comparison of Different Modes of Operation

There is approximately 100 times increase in sensitivity in preforming the MAC technology in the competitive mode as compared to the displacement mode. For example, we have completed the MAC for cocaine and biotin. Cocaine was developed in the displacement mode and showed a sensitivity of approximately 1 μg/mL. On the other hand, biotin was developed in the competitive mode and, under favorable conditions, could reach sensitivities of 10 ng/mL. For testing of solid samples, such as seized, suspected drugs of abuse, the sensitivity displayed by the displacement mode is more than adequate. For testing of substances in a biological matrix, the competitive mode must be used.

The advantages of the displacement mode is the simplicity of manufacture and its disadvantage is its sensitivity. The most difficult step in manufacture of the MAC is the preparation of the enzyme-conjugated antibody. In the displacement mode, routine steps can be employed to prepare this vital component. Once the bound antigen is prepared, it is merely soaked in the enzyme-conjugated antibody solution and washed. Any unbound enzyme or inactivated antibodies are removed at this stage of manufacture since they do not bind to the immobilized antigen. Conversely, for the competitive mode, the enzyme-conjugated antibody must be purified to remove the unconjugated enzyme and inactive antibody. If present, they would not bind to the immobilized antigen and would produce a background color, thus reducing the sensitivity of the MAC.

ADVANTAGES OF THE MAC

The MAC has a number of advantages. They include:

- The assay is rapid and sensitive. It is completed in under 5 minutes with < 10 ng/mL sensitivity.
- The presence of a drug is indicated by the development of a color.
- No instrumentation is needed.
- Antibodies are used for increased specificity over chemical tests currently used.
- The MAC is appropriately designed for field use. Only one solution is needed. (As an option, the MAC may be configured so that only water may be used.)
- Multiple assays may be incorporated on the same slide by addition of more wells.
- Controls are included to distinguish false positives, false negatives and improper storage conditions. However, they may be eliminated, if desired, after appropriate field evaluation.
- No toxic or corrosive reagents are used.

CONCLUSIONS

Development of the MAC assay has been completed for cocaine and biotin, a model compound. Since only a change in antibody and antigen is needed, other substances can easily be incorporated. The sensitivity of the assay varies upon the mode used, which is determined during the manufacture of the MAC. For the displacement mode, the MAC can detect less than 1 μg/mL of cocaine using 50 μL of sample within 5 minutes. For the competitive mode, the MAC can detect less than 10 ng/mL of biotin using a 50 μl sample. The system is portable and has an estimated shelf-life of over one year. The MAC assay is flexible enough to be configured to the requirements of the client.

REFERENCES

1. D.A. Kidwell, "Superabsorbent Polymers - Media for the Enzymatic Detection of Ethyl Alcohol in Urine", *Anal. Biochem.*, **182** 257-261(1989).

2. H.G. Curme, R.L. Columbus, G.M. Dappen, T.W. Elder, W.D. Fellows, J. Figueras, C.P. Glover, C.A. Goffe, D.E. Hill, W.H. Lawton, E.J. Muka, J.E. Pinney, R.N. Rand, K.J. Sanford, and T.W. Wu, *Clin. Chem.* **24** 1335-1342(1978).

N92-22698

DETECTION OF SMALL MOLECULES WITH A FLOW IMMUNOSENSOR

Anne W. Kusterbeck & Frances S. Ligler
Center for Bio/Molecular Science and Engineering,
Naval Research Laboratory,
4555 Overlook Avenue, Washington, DC 20375-5000

ABSTRACT

We describe the development of an easy-to-use sensor with widespread applications for detecting small molecules. The flow immunosensor can analyze discrete samples in under 1 minute or continuously monitor a flowing stream for the presence of specific analytes. This detection system is extremely specific, and achieves a level of sensitivity which meets or exceeds the detection limits reported for rival assays. Because the system is also compact, transportable, and automated, it has the potential to impact diverse areas. For example, the flow immunosensor has successfully detected drugs of abuse and explosives, and may well address many of the needs of the environmental community with respect to continuous monitoring for pollutants. Efforts are underway to engineer a portable device for use in the field.

INTRODUCTION

Market demand for a rapid means of detecting small molecules in diverse environments has increased dramatically in recent years. Monitoring for environmental contaminants, screening for drugs of abuse, and controlling fermentation processes all require accurate and sensitive determinations of what compounds are present in a complex sample. Current methods rely on optical sensing, which does not provide specificity, or on more complex analytical techniques, which are time consuming and labor intensive. The Center for Bio/molecular Science and Engineering at the Naval Research Laboratory has developed a sensor which operates in flow and is capable of detecting small molecular weight compounds in the parts per billion range in minutes (1,2).

BACKGROUND

The flow immunosensor capitalizes on the specific recognition between an antibody and its selected target molecule, the antigen. With the current technologies, large quantities of antibody with a defined specificity can be readily produced in the laboratory. This confers a great degree of flexibility on this sensor with respect to the compounds it can be designed to detect.

Though other antibody assay systems are in use, they rely on either direct binding of antigen (i.e. sandwich immunoassays) or a competitive binding of antigen versus labeled antigen. Only the latter configuration has been widely used for the detection of small molecules. The flow immunosensor relies on a distinctly different type of antibody-antigen interaction. This device measures antigen displacement events rather than binding events.

TECHNICAL APPROACH

The key elements of the sensor are: 1) the antibody specific for the target, 2) signal molecules similar to the target, but labeled so they are highly visible to a detector, and 3) a detector. Antibodies are chemically immobilized on a solid support, for example glass beads. All the antibody binding sites are saturated with a fluorescently-labeled analog of the antigen, creating an antibody/labeled antigen complex. The support is packed into a 1/2 inch column and connected to a water stream. The target molecule is introduced via a continuous or pulsed stream under nonequilibrium conditions. Only the target molecules in a complex sample are recognized. When the target molecule is present in the sample, the labeled antigen is displaced from the antibody and moves into the flow stream. The displaced labeled antigen enters the flow cell of a simple fluorimeter and triggers a response.

The signal from the detector will be proportional to the number of target molecules in the sample. No reagent addition is required throughout the assay and analysis time is minimal. A laptop computer is used to control system hardware, as well as handle data acquisition and analysis. A schematic of the system is shown in Figure 1.

EXAMPLE OF FLOW IMMUNOSENSOR USE

The flow immunosensor developed at NRL has successfully detected both drugs of abuse and explosives present at low levels in water samples. For example, to detect cocaine, an antigen to the immobilized antibody was reacted with fluorescein-labelled benzoylecgonine, the major metabolite of cocaine. Samples containing cocaine were introduced into the water stream. Analysis time was rapid, as detection of the cocaine was accomplished within 1 minute. Multiple samples could also be detected on the same column. Figure 2 illustrates the application of 750 ng/ml samples of cocaine to a single column. Over 50 positive samples were detected.

The explosive molecule TNT has also been tested using the flow immunosensor. As shown in Figure 3, when TNT was introduced into the flow stream, the magnitude of the signal was proportional to the amount of TNT introduced into the system. The response was specific for TNT, as a compound with a similar ring structure did not generate a signal.

ADVANTAGES OF THIS APPROACH

The flow immunosensor has many advantages over existing technologies. Operation of the sensor is straightforward and fast, and does not require a skilled operator or extensive training. The prototype now in use requires only two computer keystrokes. In its simplest version, the user introduces the sample at the beginning of the system and records the results within 1 minute of sample introduction. Again, this is in contrast to the user intensive and time consuming operation of currently available detection devices. The widely used methods often require addition of different reagents throughout the assay and lengthy incubation times, or demand the use of large, sophisticated instruments. In the NRL sensor, all the components required to recognize the target and release a signal are contained within a small column.

The flow immunosensor is also well-characterized. Experimental parameters, including column size and flow rate, have been studied extensively. Using equations derived at NRL, we are able to predict the behavior of the sensor for a given antibody-analyte pair. In addition, because the immunosensor is antibody-based, detection is extremely specific for the target molecule.

System manufacturing costs and portability are also important considerations. The components of the current system are inexpensive and off-the-shelf. Cost for the laboratory prototype is under $10,000, and the sensor can be engineered to fit into a single briefcase with microprocessor control.

An additional strength of the NRL detector is that it can be used in a variety of environments. It is readily adaptable for use with individual samples injected by hand, air samplers that extract vapors into water, or super sipper systems that rapidly inject samples from hundreds of vials.

Finally, the detection limit of the flow immunosensor is already comparable to established, more complicated systems. Using the NRL sensor, cocaine and TNT in water have been detected at levels of less than 5 parts per billion (equivalent to 5 ng/ml). This level of sensitivity is well-below that obtained using precipitation, dip stick, ELISA and fluorescence polarization methods, and is comparable to radioimmunoassays.

CONCLUSION

A flow immunosensor has been developed and tested at NRL which will successfully detect both drugs of abuse and explosives present at low levels in water samples. This detection system relies on the ability of sample antigen to displace labeled antigen from antibody under flow conditions. The laboratory prototype is currently being used to detect cocaine, heroin, and TNT. The advantages of the NRL sensor include its speed, specificity, and versatility with regard to measurement of a range of analytes and sampling conditions.

NOTES AND ACKNOWLEDGEMENTS

Two patents (3,4) have been filed covering this technology and are available for license. This work was supported by the Federal Aviation Administration and the U.S. Customs Service. The authors thank Reinhard Bredehorst, Robert Ogert, Paul Charles, Jeffrey Chilla, and James Whelan for their scientific contributions. The views expressed here are the authors' own and do not reflect the policy of the U.S. Navy, Department of Defense or United States Government.

REFERENCES

1. A. W. Kusterbeck, G. A. Wemhoff, P. Charles, R. Bredehorst, F. S. Ligler, A Continuous Flow Immunoassay for Rapid and Sensitive Detection of Small Molecules, *J Immuno Meth* 135, 191-197 (1990).

2. Bredehorst, R., Wemhoff, G. A., Kusterbeck, W. E., Charles, P. T., Thompson, R. B., Ligler, F. S., & Vogel, C.-W., Novel Trifunctional Carrier Molecule for the Fluorescent Labeling of Haptens. *Anal Biochem* 193, 272-279 (1991).

3. Ligand-Label Conjugates which Contain Polyoxoanions of Sulfur or Phosphorus, (submitted 4/90).

4. Flow Immunosensor Method and Apparatus, Patent Application # 07-486024 (submitted 2/90)

FIGURES

Figure 1. Flow immunosensor operation
Figure 2. Repetitive detection of cocaine samples.
Figure 3. Detection of TNT with the flow immunosensor

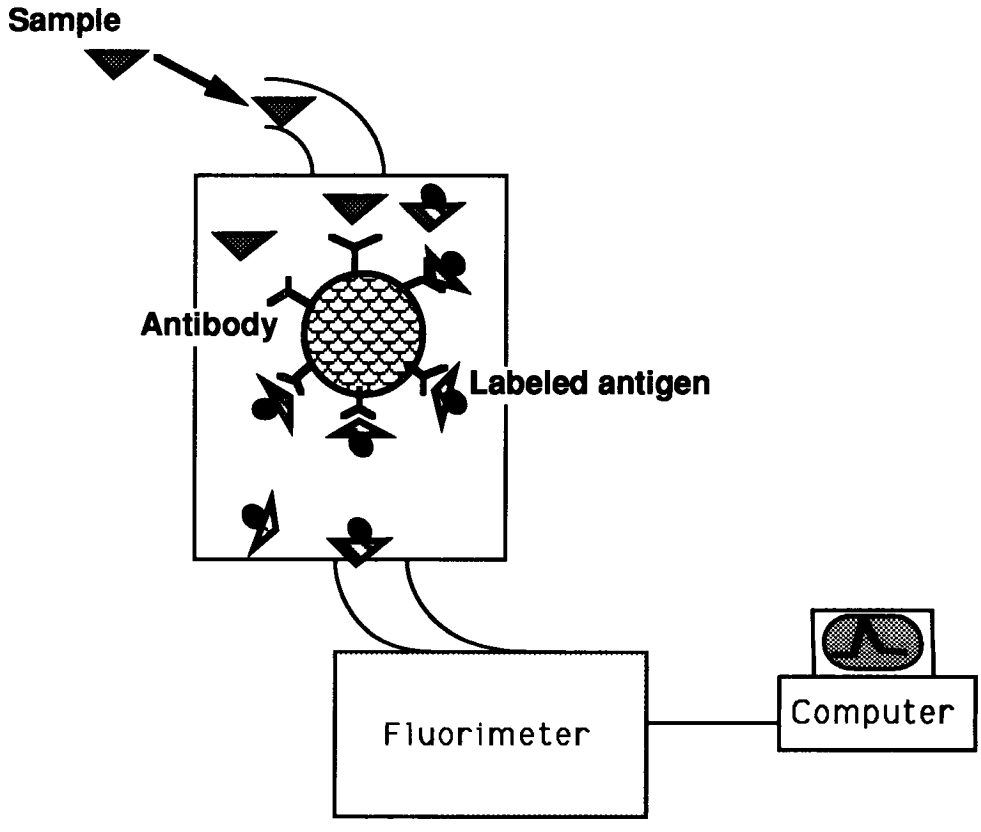


Figure 1 - Flow immunosensor operation. A sample is added to a small column containing immobilized antibodies saturated with labeled antigen molecules. A proportional amount of labeled antigen is displaced by the sample and is moved by the flow stream to the fluorimeter. Signal output from the fluorimeter is transmitted to the computer and analyzed.

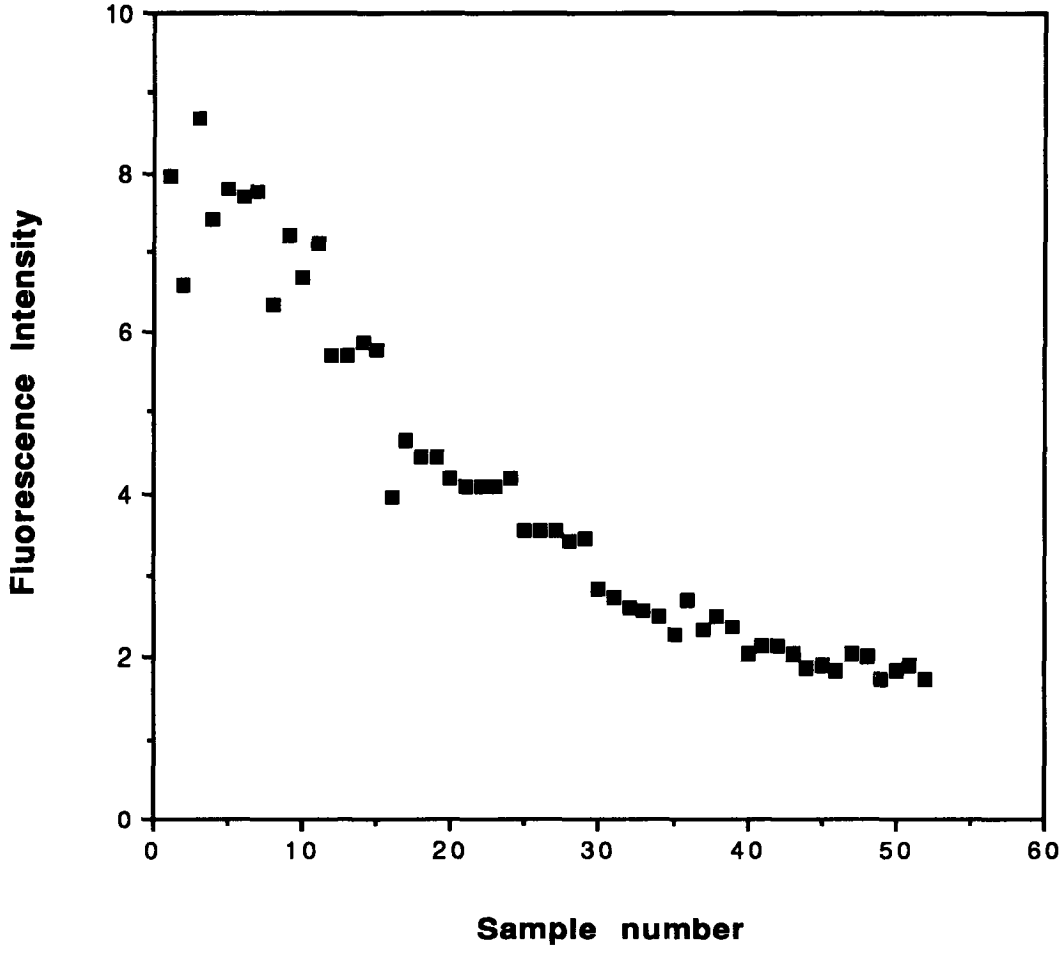


Figure 2 Repetitive detection of cocaine samples. Consecutive samples of cocaine, 200 µl at 750 ng/ml, were applied to the flow immunosensor. The integrated peak area of fluorescence intensity was calculated for each sample and plotted.

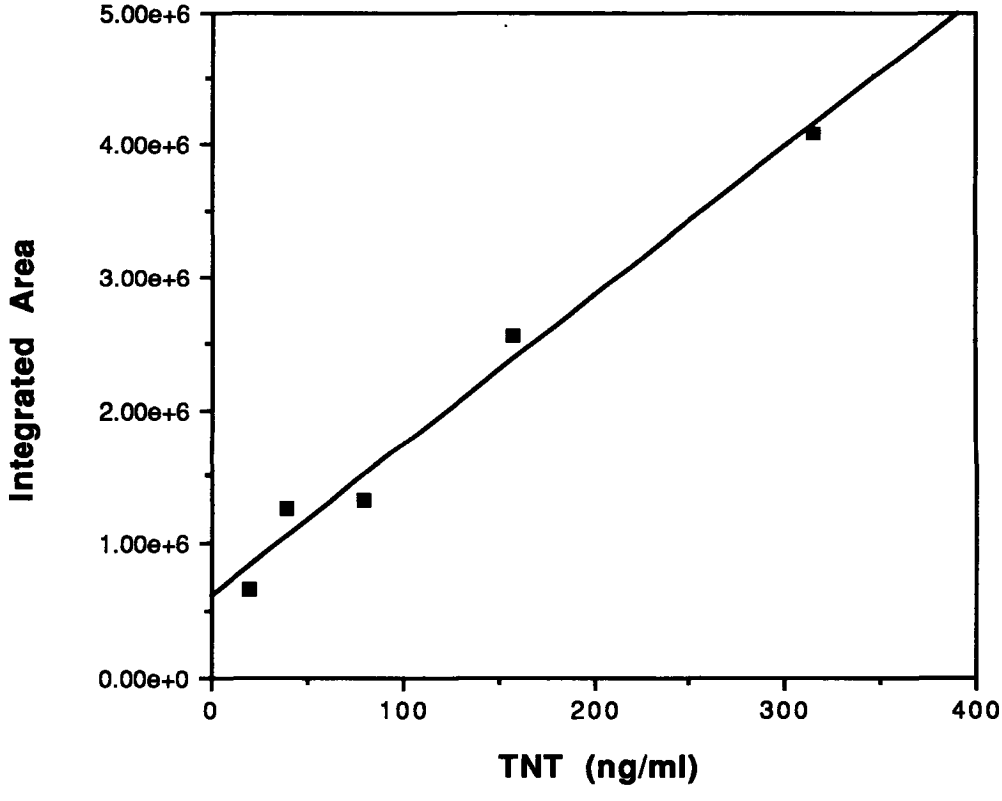


Figure 3 Detection of TNT with the flow immunosensor

N92-22699

NUCLEIC ACID PROBES IN DIAGNOSTIC MEDICINE

Phillip A. O'Berry
National Technology Transfer Coordinator for Animal Science
National Soil Tilth Laboratory, USDA, ARS, OCI
2150 Pammel Drive, Ames, Iowa 50011

ABSTRACT

The need for improved diagnostic procedures will be outlined and variations in probe technology will briefly be reviewed. A discussion of the application of probe technology to the diagnosis of disease in animals and humans will be presented. A comparison of probe vs. nonprobe diagnostics and isotopic vs. nonisotopic probes will be made and the current state of sequence amplification will be described. The current market status of nucleic acid probes will be reviewed with respect to their diagnostic application in human and veterinary medicine. Representative product examples will be described and information on probes being developed that offer promise as future products will be presented.

INTRODUCTION

New technology will be required to meet the exponentially increasing world food needs. World population is projected to double in the four decade period from 1960-2000 to reach 6.1 billion [1]. World livestock and poultry numbers have also increased significantly. Healthy animals contribute immeasurably to the nutritional status, economic productivity, land resource utilization and overall well-being of our global society. Molecular approaches to agricultural production and animal disease diagnosis provide a ray of hope that world food needs can be met. Nucleic acid probes provide a powerful new technology that is just now being applied to improve our ability to more rapidly detect diseases and their causative agents.

Diagnostic Technology

Diagnosis of the hundreds of infectious and noninfectious diseases that affect man and animals is not simple. It often involves clinical examination, serologic tests, cultural or microscopic examination of a wide range of samples, fluids, or tissues, postmortem examination and the gathering and assessment of additional information. Adequate diagnostic tests do not exist for many important diseases. Even in instances where adequate tests do exist, there is a critical need for an improved test that is more rapid, more economical or more reliable. Thus, the field of diagnostics lends itself to the application of nucleic acid probe technology.

In simplest form, a nucleic acid probe is a specific piece of single stranded nucleic acid, usually composed of highly conserved nucleotide sequences, that contains an isotopic or nonisotopic label allowing it to be visualized when it finds and binds (hybridizes) to another piece of nucleic acid that has complementary nucleotide bases. It has been known since 1953 that the structure of DNA is a double helix [2]. It was later found that two separate but complementary strands of DNA reassociate (hybridize) under specific conditions to form a double helix [3]. These and other principles were used in the late 1960's to identify specific pieces of DNA in the nuclei of cells fixed to a microscope slide using labelled complementary pieces of nucleic acid. Thus probe technology was born.

Probes are made from such materials as target-specific sequences of cloned genes, complementary DNA, natural or synthetic single-stranded oligonucleotides or ribosomal RNA. Sample material usually is solubilized, and either its proteins are electrophoretically separated and bound to an appropriate support membrane (Western blotting) or they are applied to nitrocellulose filters (Dot blotting) for reaction with a probe. The probe, labeled by incorporation of a radioisotope (e.g. ^{32}P or ^{125}I) or enzyme, (e.g. alkaline phosphatase or horseradish peroxidase) then searches the sample on the support system for homologous nucleotide sequences, or "target" DNA. Probe-target DNA hybrids are then visualized or enumerated by autoradiographic or direct scintillation counting or by enzyme-substrate color change reactions.

Most diagnostic probes developed thusfar utilize radioisotopes as labels with resulting advantages and disadvantages. The advantages are high sensitivity, easy signal detection and permanent record of test results. The disadvantages are the risk of exposure to harmful levels of radiation, expensive safety protection equipment, strict waste disposal procedures and relatively limited probe storage life. The diagnostic community has been slow to develop, market and use these radioactive probes.

Nonisotopic probes offer some distinct advantages over their isotopic counterparts. They are more simple to run, obviously pose no radiation hazard and require no specialized safety equipment or disposal procedures. These probes tend to have longer storage life. The recent development and marketing of chemiluminescent substrates for use with nonisotopic probes allow for exposure of x-ray film in the probe reaction with the production of a hard-copy of the test results in only a few minutes. These and other recent improvements in probe development technology have made nonisotopic probes more sensitive and practical than ever before.

Powerful new procedures to amplify target DNA sequences in sample material offer considerable promise in enhancing the application of probe technology to diagnostic veterinary medicine. The best known of these is the polymerase chain reaction, (PCR) formally introduced in 1986 [4] and patented in 1987. A recent review article [5] describes PCR in detail and discusses its application to probe technology. Target DNA is heat denatured and specific paired primers are repeatedly annealed and extended with a specific polymerase until amplification is sufficient to allow signal detection even in a complex mixture. The relatively simple procedure uses a single thermostable enzyme, can be performed in a few hours and in a single tube, and lends itself to automation. Among the problems in the procedure are nonspecific hybridization, the number of steps involved in the thermal cycling of the reaction, the number, and sometimes efficiency, of product molecules per cycle is rather low thus necessitating an increased number of cycles to reach a sufficient amplification level. In addition, RNA targets and DNA targets are indistinguishable in some mixed nucleic acid samples. However, new developments in PCR techniques allow amplifications of over 100 billion-fold, meaning that a single molecule can be amplified to a point of visibility on a simple agarose gel.

Several other target amplification procedures are also available. They are amplification by RNA transcription (TAS, 3SR) and amplification by ligation (LAR/LAS). The comparative advantages and disadvantages of these procedures and PCR are well described. It is likely that newer and even more useful procedures will come in the years ahead.

Applications of Probe Technology

Nucleic acid probes are uniquely and specifically well suited to detect even minute amounts of their homologous counterparts. Applications of this powerful new technology are rapidly increasing and virtually limitless.

Detection of disease agent nucleic acids in diagnostic samples is an obvious application. Molecular approaches to infectious disease diagnosis has been recently reviewed [6]. Systems must be specifically developed for use on such varied diagnostic samples as tissues, cells, blood, urine, saliva, feces, mucosal swabs, skin scrapings, semen or other specimen. All must be compared against and be more advantageous than current conventional cultural, histopathologic, microscopic or immunologic diagnostic methods.

Nucleic acid probes can also aid in the differentiation between closely related microorganisms. Detection of the presence of a specific serovar or strain can influence the therapeutic or preventive measures that should be taken. Clinical samples are often complicated by the presence of nucleic acids of closely related nonpathogens or pathogens of such low virulence as to be unimportant.

The typing of DNA using probes has many practical applications. They include, but are not limited to, the detection of genetic defects, genome mapping, forensic uses, sex determination in sperm cells, paternity determination and microbial classification.

Some applications of nucleic acid probe methodology have been made to the detection and prognosis of certain malignancies. Some conditions have been associated with viral infections, such as cervical cancer and human papilloma. Others, such as certain breast and ovarian carcinomas, have been associated with oncogene amplification. Certain leukemias and lymphomas have been connected to specific clonal gene arrangements and chronic myelogenous leukemia has been shown by DNA probe use to be a specific gene translocation. Thus nucleic acid probes offer considerable promise as malignancy diagnostic, prognostic and perhaps even therapeutic applications.

An additional application of nucleic acid probe technology is in the area of quality control of biologics. Probes represent a powerful new tool to screen for and detect adventitious microorganisms in biologicals. Certain microorganisms are commonly found in cell cultures or production materials, like serum, and are quite difficult to detect by current and more conventional methods.

At this point in the development of nucleic acid probe technology, the most common application is in research. Probes are powerful research tools used for a wide variety of nondiagnostic purposes. They have contributed significantly to the gathering of new information on pathogenesis, epidemiology, latency, microbial genetics, taxonomy and microbial ecology. Probes are also used to develop new and improved immunogens, diagnostic tests and/or therapeutic agents.

Probe Products in Human Diagnostics

The market was surveyed to determine the number and types of nucleic acid probes available for diagnosis of human diseases. All companies in a recent product listing [7] were contacted, literature was surveyed and numerous contacts were made with Food and Drug Administration officials.

Over 200 DNA or RNA probes are currently commercially available. These products are marketed by 23 companies and the American Type Culture Collection with most specified for research purposes only. They cover such diverse areas as viruses (Human Immunodeficiency Virus, Human Papillomavirus, Herpes Simplex Virus, Epstein-Barr Virus and Hepatitis A and B viruses), bacteria (Legionella, Mycobacteria, Campylobacter, Mycoplasma and Haemophilus), mycotic agents (Histoplasma, Blastomyces and Coccidioides), drugs (digoxin and phenobarbital), neoplasias (Chronic Myelogenous Leukemia, Acute Lymphocytic Leukemia and certain Lymphomas), genetic diseases (Cystic Fibrosis and Duchene Muscular Dystrophy) as well as a host of oncogenes, proto-oncogenes, all 24 human chromosomes and over 100 probes for specific human Restriction Fragment Length Polymorphisms.

Nucleic acid diagnostic probes are considered to be devices by the FDA and as such are required to receive premarket approval if marketed for diagnostic use. None are licensed by FDA but 61 have received FDA approval in the past 6 years. One company, Gen-Probe Inc., San Diego, has received approval for 30 probes. Probes marketed for research purposes only receive limited FDA oversight.

It is difficult to compile information on probe product sales revenues. However, two companies active in this area (Enzo Biochem, Inc. and Oncogene Science) reported gross sales from all products in 1990 of $19 million and $5 million respectively [8]. It is not known how much of these gross sales were derived from probe sales. The market for probes in the second half of this decade has been estimated to be $500 million [9].

Probe development and use for human diagnostics has grown exponentially in the last few years. It is likely that many new and perhaps more powerful probes will be developed in the very near future. Promising new probes are being developed for Cryptosporidia [10], Borrelia burgdorferi [11], neoplastic tissue [12], E. coli in foods [13], Salmonella in foods [14], Listeria in foods [15], and Human Papillomavirus in Pap smears [16]. Ultimately, the value of these tests and any other probe tests that reach the market will be determined in the same way as for all new diagnostic products - by independent scrutiny of their clinical performance, convenience and relative cost [17].

Probe Products in Veterinary Diagnostics

A similar market survey was conducted to determine the number and types of nucleic acid probes available for diagnosis of animal diseases. Literature and scientific publications were surveyed and contacts were made with U.S. Department of Agriculture officials in the Animal and Plant Health Inspection Service.

Over 100 probes developed against a wide range of animal viral and bacterial pathogens have been reported [18]. Although premarket approval is not mandatory, one probe has received USDA approval to be marketed for veterinary diagnostic purposes. It is a DNA probe developed by IDEXX Corporation of Portland, ME for the detection of Mycobacterium paratuberculosis in fecal samples. A second probe, developed by scientists at Montana State University is used by Agricultural Bioengineering, Inc. of Bozeman, MT to test samples submitted to it for Tritrichomonas foetus, the flagellated protozoan that causes the reproductive disease bovine trichomoniasis. Little information is available on these probe product sales since both were introduced this year.

The future looks promising that additional probes will enter the veterinary diagnostics market. They will cover an expanding number of pathogens and diseases. Many will have patent protection and exclusive license provisions. Several of the more promising probes are specific for such diverse organisms as Treponema hyodysenteriae [19], Campylobacter fetus and Campylobacter hyointestinalis [20], Anaplasma marginale [21], Bovine Virus Diarrhea Virus [22], Eperythrozoon suis [23] and Bovine Leukocyte Adhesion Deficiency [24], a genetic disease of cattle.

SUMMARY

A powerful new technology utilizing labeled nucleic acid is being applied to human and veterinary diagnostics. Over 300 nucleic acid probes have been developed in the past few years. This new technology will be increasingly transferred into the U.S. and global marketplace with considerable promise of improvement in human and animal health.

REFERENCES

1. Pritchard, W. R. JAVMA 198, No. 1, Jan. 1, 1991: 46-51.
2. Watson, J. D. and Crick, H. C. Nature 171, 1953: 964-967.
3. Marmur, J. and Doty, P. J. Molec. Biol., 3, 1961: 595-617.
4. Mullis, K. Faloona, F. Scharf, S., Saiki, R., Horn, G. and Erlich, H. Cold Spring Harbor Symposium on Quantitative Biology 51, 1986: 263.
5. Arnhelm, N. and Levenson, C. H. Chem. & Eng. News, Oct. 1, 1990: 36-47.
6. Charache, P. The Fifth San Diego Conference. Nucleic Acids: New Frontiers. Nov. 14-16, 1990.
7. Scherago, E. J. Science 247, Part II, March 23, 1990: G83-92.
8. Thayer, A. M. Chem. & Eng. News, April 1, 1991: 21.
9. Parr, R. S. Biotech Patent News 5, No. 3, March 1991: 4.
10. Turner, M. H. The Scientist 4, No. 21, Oct. 29, 1990: 4-7.
11. Goodman, J. L., Jurkovich, P., Kramber, J. M. and Johnson, R. C. Infection and Immunity 59, No. 1, Jan. 1991: 269-278.
12. Parr, R. S. Biotech Patent News 5, No. 2, Feb. 1991: 8.
13. Hsu, H. Y., Chan, S. W. Sobell, D. I., Halbert, D. N. and Groody, E. P. J. Food Protection 54, No. 4, April 1991: 249-255.
14. Wilson, S. G., Chan, S. W., Deroo, M., Vera-Garcia, M., Johnson, A., Lane, D. and Halbert, D. N. J. Food Protection 55, 1990: 1394-1398.
15. King, W., Raposa, S., Warshaw, J., Johnson, A., Halbert, D. and Klinger, J. D. Int. J. Food Microbiol. 8, 1989: 225-232.
16. Parr, R. S. Biotech Patent News 5, No. 3, March 1991: 4.
17. Engleberg, N. C. ASM News 57, No 4, 1991: 183-186.

18. Knowles, D. P. and Gorham, J. R. Rev. Sci. Tech. Off. Int. Epiz. 9, No. 3, 1990: 733-757.
19. Jensen, N. S., Stanton, T. B. and Casey, T. A. U.S. Patent & Trademark Office Serial No. 07/496,579, March 21, 1990.
20. Wesley, I. U.S. Patent & Trademark Office Serial No. 07/603,503, Oct. 26, 1990.
21. Goff, W. L., Stiller, D., Roeder, R. A., Johnson, L. W., Falk, D., Gorham, J. R. and McGuire, T. C. Vet. Microbiol. 24, 1990: 381-390.
22. Kelling, C. L., Kennedy, J. E., Rump, K. K., Stine, L. C., Paul, P.S. and Partridge, J. E. Am. J. Vet. Res. 52, No. 8, Aug. 1991: 1237-1244.
23. Oberst, R. D., Hall, S. M., Jasso, R. A., Arndt, T. and Wen, L. Am. J. Vet. Res. 51, No. 11, Nov. 1990: 1760-1764.
24. Kehrli, M. E. and Shuster, D. U.S. Patent & Trademark Office, Sept. 20, 1991.

N92-22700

THE ROTATING SPECTROMETER: BIOTECHNOLOGY FOR CELL SEPARATIONS

David A. Noever
NASA Marshall Space Flight Center
Huntsville, AL 35812

ABSTRACT

A new instrument for biochemical studies, called the rotating spectrometer, is shown to separate previously inseparable active cell cultures. The rotating spectrometer is intended for use in pharmacological studies which require fractional splitting of heterogeneous cell cultures based on either cell morphology or swimming behavior. As a method to separate and concentrate cells in free solution, the rotating method requires active organism participation and can effectively split the large class of organisms known to form spontaneous patterns. Examples include the biochemical 'star', an organism called *Tetrahymena pyriformis.* Following focusing in a rotated frame, the separation is accomplished using different radial dependencies of concentrated algal and protozoan species. The focusing itself appears as concentric rings and arises from the coupling between swimming direction and Coriolis forces. A dense cut is taken at varying radii and extraction is replenished at an inlet. Unlike standard separation and concentrating techniques such as filtration or centrifugation, the instrument is able to separate motile from immotile fractions. For a single pass, typical split efficiencies can reach 200-300% compared to the inlet concentration. Applications for pharmacological studies are discussed.

INTRODUCTION

Most biological concentration techniques treat the cells passively, relying on their density for centrifugation or size for filtration. The recent advent of a gyrotactic spectrometer has introduced new ways to separate organisms based on more subtle specifications, such as swimming speed or morphology [1]. However, even this technique is known to focus gyrotactic organisms only--namely organisms with their bouyant center displaced from their mass center or equivalently, bottom heavy species. This new class of separation has not proven successful for separating many organisms of biochemical significance [2], most notably a "biochemical star" protozoa called *Tetrahymena.* This organism, in particular, has numerous applications in drug testing and vitamin assays [3]; its biochemistry corresponds remarkably to that found in humans and rats. A recent review [4] remarked that "for more than four decades it has been the organism of choice in analyses such as in evaluation of protein quality, determination of effects of carcinogens, insecticides, fungicides, mycotoxins, oil product and organic chemicals, as well as antimetabolites, heavy metals and pharmaceutical drugs."

The rotating spectrometer described herein (Fig. 1) accomplishes separation/ concentration using radial dependencies of different microorganisms (including *Tetrahymena).* This focusing into concentric rings

(Fig. 2) arises from the coupling between swim directions and the Coriolis force in a rotating frame. A dense cut of a particular organism is taken at various radii and this extraction is replenished at an inlet. The apparatus has been demonstrated to concentrate the following test organisms: algae (*Gonium pectorale, Euglena gracilis, Euglena gracilis* var. *bacillarus, Polytomella parva*) and protozoans (*Tetrahymena pyriformis*). This separation is accomplished independent of whether the organisms are gyrotactic.

Unit operation requires filling of a rotating chamber (Fig. 1) with organisms and solution to be separated and concentrated (at $>2 \times 10^5$ organisms/ml). This initial concentration can be accomplished by standard means of either filtration or centrifugation; both have been employed successfully as detailed in the section on data collection. After the turntable begins rotation, dense organism rings form spontaneously in 5-30 seconds. A dense cut is taken using a point scoop at the desired radii. The apparatus can be stopped for refilling or more solution can be added continuously at an inlet. This process can be repeated in conjunction with other separation techniques.

Experiments described elsewhere [5] were the first to find this new rotating bioconvective structure, namely an ordered set of concentric rings. This memorandum describes the design and performance of a rotating spectrometer which uses this ring structure to split microorganism cultures. The spectrometer design incorporates the following features: (1) it can be used with any class of organisms known to bioconvect (and not merely gyrotactic swimmers); (2) samples are easily removed and replenished owing to the device's top-loading option; (3) the selectivity of the dense cut or split can be tuned with respect to rotation speed and fluid height to take optimal test samples; (4) because all organisms are known not focus to a centerline (as in previous gyrotactic spectrometers), several cuts can be taken at different radii that show different organism properties; and 5) small (< 4 cm^3) and large (> 100 cm^3) samples can be separated or concentrated. Small sample requirements can prove advantageous, particularly if measurements incorporate several process steps which one wishes to monitor, but not subject the entire sample to an intrusive concentrative scheme. It is worth noting that in the limit of dead cells, cultures do not separate or concentrate in a rotating frame.

Tuning of a separation can be accomplished in many ways. When coupled to the rotating turntable, a voltage reducer varies angular speed continuously from rest to maximum (in the present case, 4.7 rad s^{-1}). Fluid depth can be controlled by either top-loading or inlet flow regulation. Finally, the time of sample taking can be varied, such that the concentrated stream is more or less diffuse.

SAMPLE PREPARATION AND SPECTROMETER CONSTRUCTION

Theory of Bioconvective Rings

Bioconvection is a fluid instability akin in some respects to thermal Benard cells in appearance, but driven by microorganisms' metabolic power, not heat. Bioconvective patterns appear in part from the density inversion (heavy over light fluids) which arises naturally from heavy organisms (density, ~1.05 gm cm^{-3}) which swim upward against gravity (negatively geotaxic). An additional feature of some algal patterns is a coupled locomotion called gyrotaxis, or the tendency of bottom heavy organisms to focus

spontaneously into falling streams. This effect, although not universal to all bioconvecting organisms, is known to govern some algal patterns via gravity torques exercised in a velocity gradient. A feature not apparent from Fig. 2 is the dynamic nature of a bioconvective instability. Unlike other forms of biological aggregation, the characteristic shape of biconvective patterns--alternating high and low density waves--does not depend on organism cohesion; rather patterns evolve dynamically from competing gravity, diffusion, and upward swimming. Several publications [1-2] have addressed these mechanisms and only the unique features of rotating bioconvection will be dealt with here.

Jeffreys [6] first noted the stabilizing influence of rotation on thermal convection, a result explainable in subsequent theoretical work (see, e.g., Chandresakhar [7]) using the Proudman-Taylor theorem. For sufficiently fast rotation, this theorem constrains an inviscid fluid to move primarily in the plane perpendicular to the rotation axis. In classical thermal convection, this constraining plane is horizontal (gravity directed normally to unstable fluid layers), such that rotation restricts vertical energy exchange. To counteract this limited energy exchange, however, narrow viscous channels evolve as exceptions to the Proudman-Taylor constraint against vertical transport. In unstable solutions of microorganisms, these viscous channels appear as the patterns shown in Fig. 2.

The experimental parameters of interest can be classified using a bioconvective Rayleigh number

$$Ra = g(\Delta\rho/\rho)h^2/ U\nu, \quad (1)$$

which is the dimensionless buoyancy of organisms heavier than their suspending solution, the Taylor number,

$$Ta = 4\Omega^2 h^4/\nu^2, \quad (2)$$

which is the dimensionless angular speed of rotation, where g is the gravitational acceleration, $(\Delta\rho/\rho)$ the surface density magnification (~0.10), h the fluid depth, U the swimming velocity (~0.1 cm s^{-1}), ν the kinematic viscosity (~0.01 cm^2 s^{-1}), and Ω the angular velocity of the rotating layer. One further scaling relation can be borrowed from work on thermal convection, namely the Ekman length,

$$\lambda = (\nu/\Omega)^{1/2}, \quad (3)$$

which gives a measure of the characteristic distance between rings.

Design hardware

As illustrated in Fig. 1 schematic, the spectrometer consists of a low-aspect ratio cylindrical chamber or vessel with associated electronics and scoop for rotation and separation. The total weight of the prototype is under 10 kg with dimensions of 30 x 30 x 35 cm^3.

As suspended in their growth media, the concentrated organisms are poured into the mounted vessel. A rigid top for each vessel was removable to allow easy exchange of samples and depth variation. Excepting several qualitative observations, all data were collected using an air free-surface. For very shallow samples at high rotation, the free-surface slope was noticeable, although at these depths no bioconvective patterns formed and hence perturbations became of no consequence to actual observations. In all cases, the vessel bottom was flat and centered on a turntable using a high precision (within 5 x 10^{-4} cm) machinist's deflection instrument. The actual sample rested on a larger transparent platform or stage which allowed

transmission of backlighting (from two 45^{o} angled collimated beams). Backlit samples were photographed against a black backdrop using two remotely cooled, incandescent beams. If observations of *Euglena, Gonium,* or other strongly photosensitive species were planned, samples were photographed against a white background under normal room illumination (approx. 7 lux).

The turntable itself was belt-driven, the angular velocity of which could be varied continuously from 0.5 to 4.71 s^{-1} using a voltage reducer. Between these velocities, the Ekman length was always at least one order of magnitude smaller than the smallest vessel. The other parameter of interest, fluid depth, was varied between 0.6 and 1.9 cm for *Tetrahymena* and 0.5 and 0.8 cm for *Polytomella.* To some degree a limited height range reflects the narrowness in which strong patterns form actively, although the present work did not attempt to span entirely the critical parameters. Even so, for *Tetrahymena pyriformis* (*Polytomella parva*,respectively), experimental conditions yielded bioconvective Rayleigh numbers between $Ra = 2.21 \times 10^{5}$ and 3.54×10^{5} ($Ra = 1.22 \times 10^{4}$ and 3.14×10^{5}) and Taylor numbers between $Ta = 2.47 \times 10^{6}$ and 11.56×10^{6} ($Ta = 1 \times 10^{5}$ and 4×10^{5}).

Sample splits were taken using visual alignment of pre-evacuated glass pipettes at varying radii. When lowered along a bioconvective ring, the pipette pressure was released manually and a sample was taken. Sample volume could be varied between 0.1 ml to several ml in the smallest rotated chamber (4.8 cm diameter, 1 cm height). No changes in material or chamber dimension were required for separating different species. The pre-loaded (and unconcentrated) microorganisms were cultured in their growth media [8] in a class 30K clean room held at a constant 20^{o} C and exposed to (7 lux intensity) white fluorescents. A given harvest was filtered using a 0.22 micron mesh to between 1.5×10^{5} and 2×10^{6} cm^{-3}, depending on the species and growth time. Hence patterns in these chambers arose as the organized behavior of approximately 10^{7} randomly swimming organisms!

PERFORMANCE

Separation Efficiency

The design of the rotating spectrometer proved capable of concentrating all bioconvecting species attempted, whether gyrotactic or not. Spin-up of the cultures was accomplished in a few seconds and within a minute, a two- to three-fold density magnification was obtained. The voltage reducer allowed optimization of angular speed, such that a given culture could be repeatedly and maximally split.

Spectrometer efficiency was evaluated by comparing motile organism concentrations between an adjacent dense and sparse cut. The concentrative efficiency could be measured using a hemacytometer after a given sample was fixed isotonically with a 5% glutaraldehyde-PBS solution and organisms were counted.

Concentrative factors are shown in Fig. 3 for the non-gyrotactic protozoa, *Tetrahymena pyriformis.* Samples were take in the smallest chamber (4.8 cm diameter) at 0.8 cm fluid depth. For comparison, Fig. 4 concentrative factors were obtained using an identical procedure, except at 0.6 cm fluid depth. The given cut was taken repeatedly from the the third concentric ring in each case at angular speeds varying between 3.5

and 4.7 rad s^{-1}.

Examination of Figs. 3 and 4 indicate that: (1) a maximum in separation efficiency occurs at an intermediate rotation; (2) rotation enhances separation potential compared to the stationary extrapolation to zero rotation; and (3) fluid depth and angular speed are primary design variables to optimize a split. The concentrative factor varies between a two to three times density split. Future publication will address the problem of morphology and swimming behavior within a given set of concentric rings.

The separation time of the spectrometer was also compared for different species [5]. In general, rings of the ciliated algae, *Polytomella parva*,formed more rapidly with decreasing height (to the -4.33 power) and decreasing rotation speed (to the -3.33 power). In general, *Tetrahymena* patterns formed more rapidly with decreasing height (to the -5/3 power) and decreasing rotation speed (to the -4/3 power). In the absence of a more extensive theoretical model, little else can be said about the origin of these statistically well-determined, inter-species relations except to note that faster rotation has less effects on the stability of bioconvective patterns in *Tetrahymena*. This may result from the ability of *Polytomella* to initiate viscous fingers throughout a culture's depth [6] and not just at the surface (as with *Tetrahymena)*, such that horizontal rotation constrains their stronger vertical movement to a more limited energy exchange. It is worth noting that as is the case for thermal convection, rotation in general tends to stabilize bioconvection and hence delay the onset of pattern formation.

Design Variations

Several variations on the standard rotating spectrometer were examined. The vessel shape and material were varied between a square or rectangular polystyrene dish (3 x 3 cm^2 and 3 x 5 cm^2, respectively) and a glass or polystyrene circular dish with diameter between 3.5 and 8 cm. Non-circular vessels do not affect the separation, except that the maximum ring diameter no longer corresponds with the vessel diameter, but with the largest inscribed circle. No separation differences were noted between glass and polystyrene vessels.

In contrast to results for thermal convection, bioconvective rings patterns were induced even in closed containers without a free surface. Preliminary observations using the green swimming algal species, *Euglena gracilis* var. *bacillaris*, revealed ring patterns in sealed circular dishes (4.7 cm diameter). This allows sterile separations if required to isolate samples from ambient contamination.

. DISCUSSION

For all bioconvecting algal and protozoan species tested, rotating separations and density enhnacemnts have been demonstrated. Performance evaluation has centered on concentrating capacity as a function of rotational speed and fluid depth in one of the protozoan known not to focus gyrotactically, *Tetrahymena*. Similar programs are planned for testing different concentrations, mixtures of species, chamber dimensions, etc.

Improvements in hardware design can be realized by coupling the separation with optical density measurements to find local concentrations *in vivo*. This concept would combine automated density

technologies, such as electronic Coulter cell counters, with the rotating spectrometer for optimal control.

REFERENCES

[1]. J.O.Kessler, Nature, 313, 218 (1985).

[2]. J.O.Kessler, Contemp. Phys. 26, 147-166 (1985).

[3]. Known sensitivity to vitamin B6, nicotonic acid, panthothenic acid thiotic acid, among others--see e.g. D.L. Hill, The Biochemistry and Physiology of Tetrahymena (Academic Press: NY, 1972).

[4] J.R. Nilsson, Europ. J. Protistol., 25, 2 (1989).

[5]. D. A. Noever, Reviews Sci. Instr. 62, 229 (1991).

[6]. H. Jeffreys, Proc. R. Soc. Lond. A. 118, 195-208 (1928).

[7]. S. Chandrasekhar, Hydrodynamic and Hydromagnetic Stability (Oxford: Clarendon, 1961).

[8]. R.C. Starr and J.A. Zeikus, J. Phycology (suppl.), 23, 39 (1987).

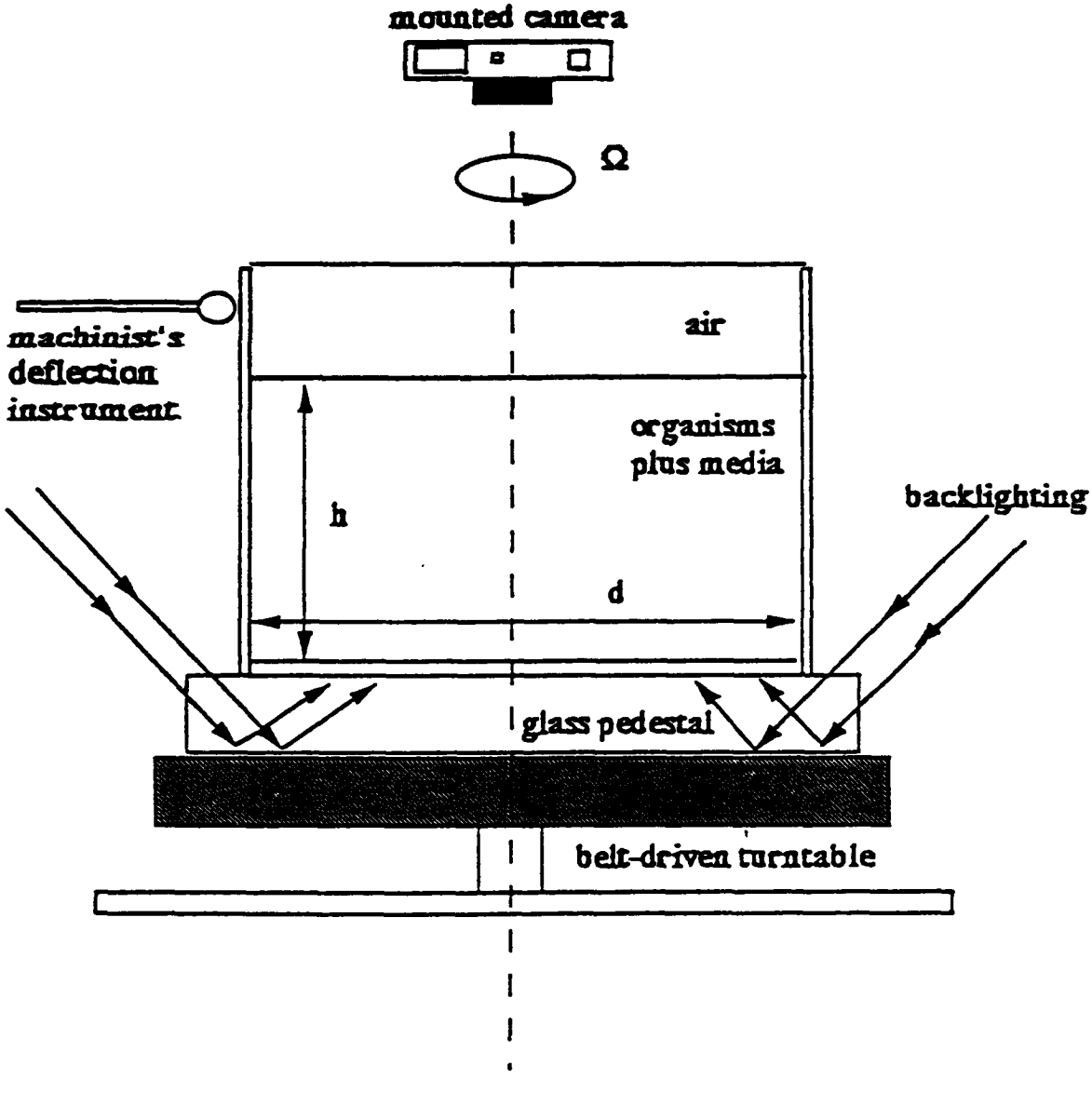


Figure 1. Top and side view of rotating spectrometer showing organism trajectories and turntable assembly used for microorganism separations and concentrations. h is the fluid depth and d is the chamber diameter.

Figure 2. Top view photograph of rotated patterns of *Tetrahymena pyriformis* in 4 mm deep culture at 4 rad/s.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

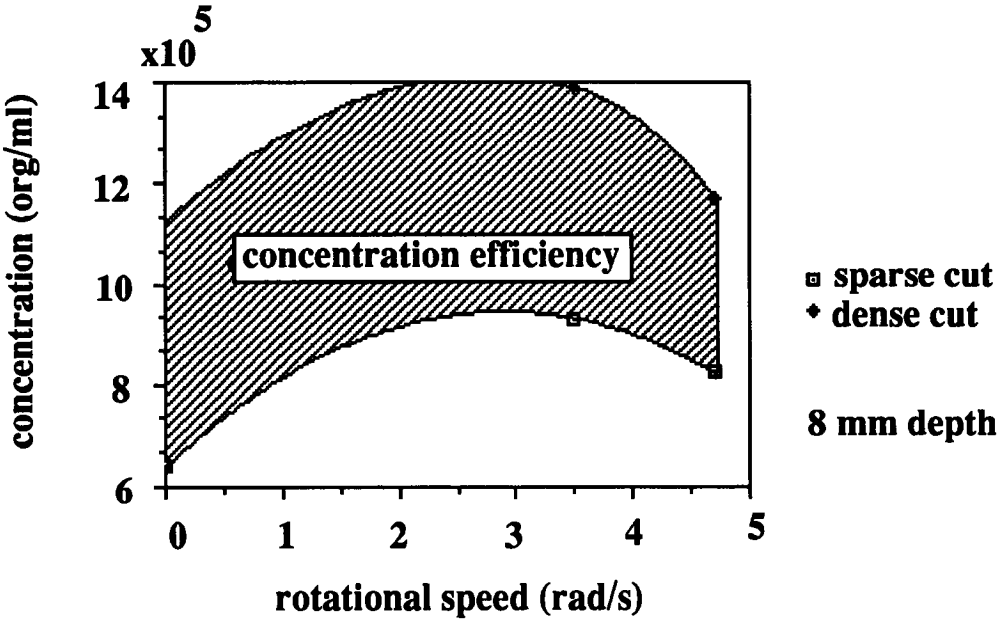


Figure 3. Concentration efficiency as a function of rotational speed at 0.8 cm fluid depth. The top line represents the dense organism cut; the bottom line shows the sparse organism cut. Concentrations are of the order of 1 million cells/ml.

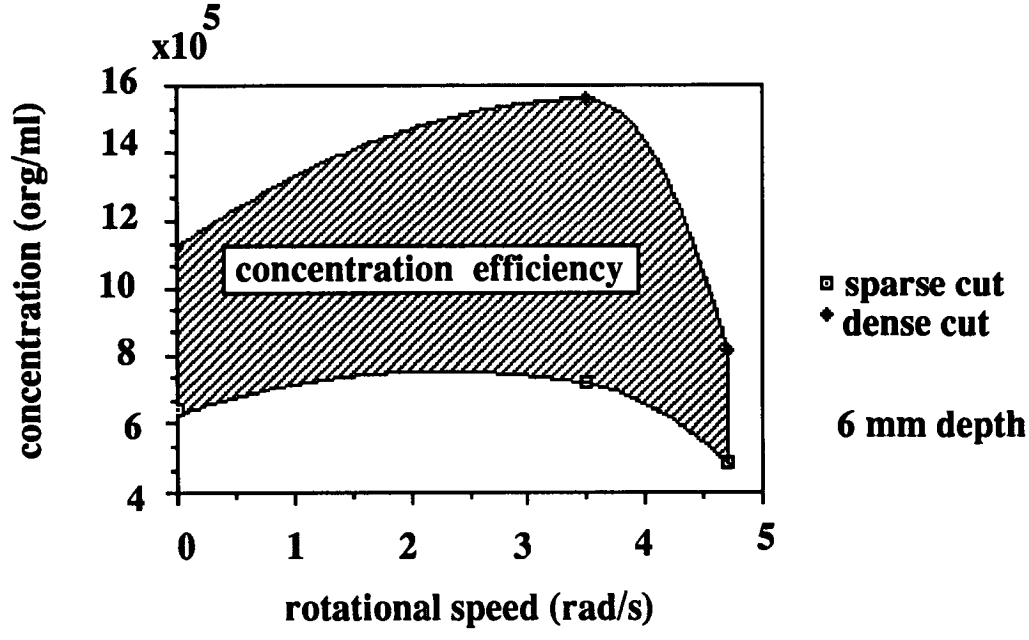


Figure 4. Concentration efficiency as a function of rotational speed at 0.6 cm fluid depth. All other variables besides depth remain constant compared to Figure 3. Notice the greater concentration factor for the shallower dishes.

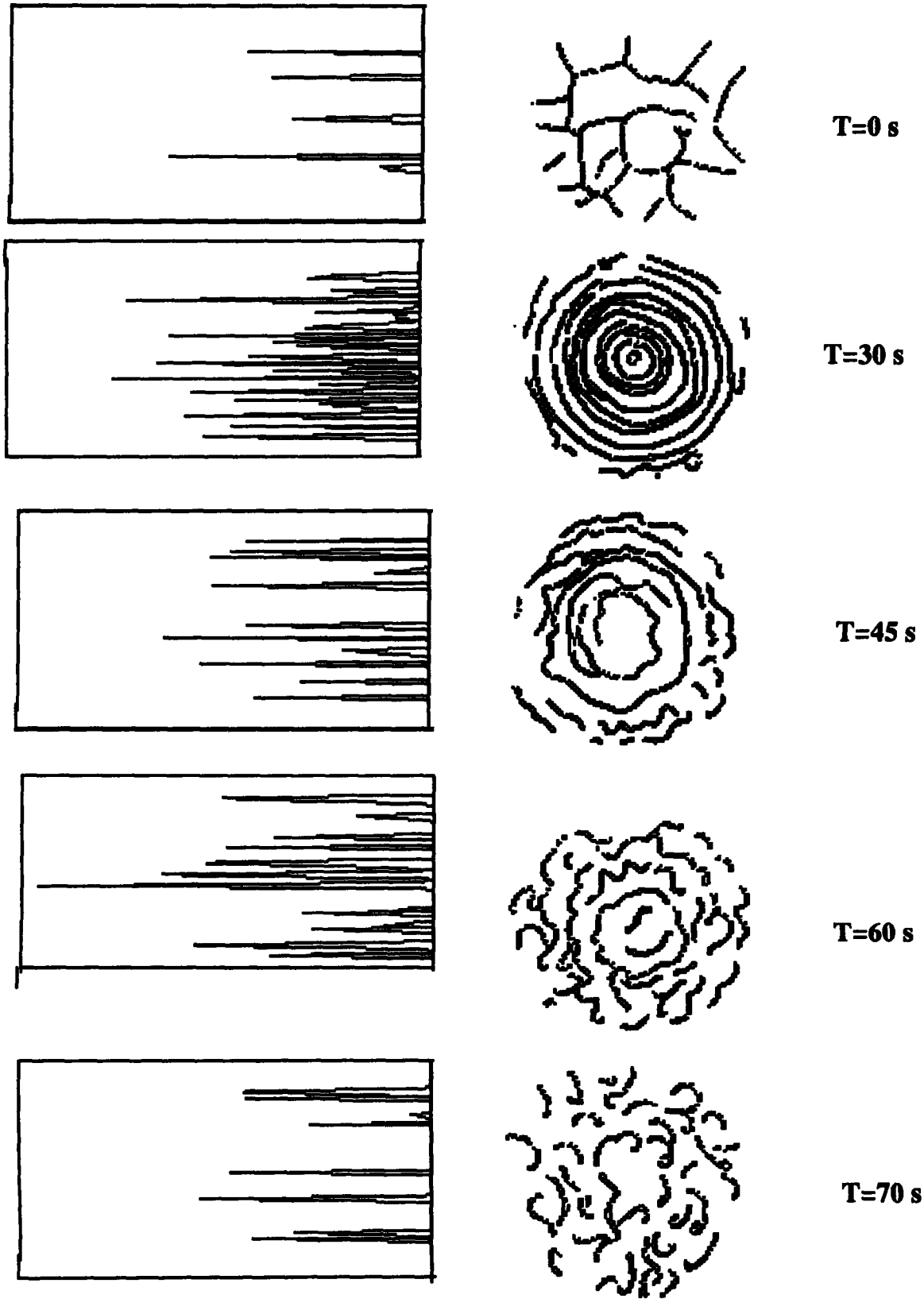


Figure 5. Progressive states of rotated patterns with increasing time for the same rotation rate. Patterns shown right, cell density along a line shown left.

ELECTRONICS

(Session D4/Room B4)

Thursday December 5, 1991

- **Method for Producing High-Quality Oxide Films on Surfaces**
- **Advanced Silicon on Insulator Technology**
- **High-Temperature Superconducting Stripline Filter**
- **An Adjustable rf Tuning Element for Microwave, Millimeter Wave, and Submillimeter Wave Circuits**

PRECEDING PAGE BLANK NOT FILMED

N92-22701

A METHOD OF PRODUCING HIGH QUALITY OXIDE AND RELATED FILMS ON SURFACES

Mark W. Ruckman and Myron Strongin
Physics Department
Brookhaven National Laboratory
Upton, New York 11973

and

Yongli Gao
Department of Physics and Astronomy
University of Rochester
Rochester, NY 14627

ABSTRACT

Aluminum oxide or aluminum nitride films were deposited on MBE grown GaAs(100) using a novel cryogenic-based reactive thin film deposition technique. The process involves the condensation of molecular oxygen, ammonia or other gases normally used for reactive thin film deposition on the substrate before the metal is deposited. The metal vapor is deposited into this layer and reacts with the molecular solid to form the desired compound or a precursor that can be thermally decomposed to generate the desired compound. The films produced by this method are free of impurities and the low temperatures can be used to control the film and interfacial structure. The process can be easily integrated with existing MBE-systems and on going research using the same appartatus suggests that photon or electron irradiation could also be used to promote the reactions needed to give the intended material.

INTRODUCTION

This paper discusses a new reactive thin film fabrication technique which has been used to deposit films like Al_2O_3 or TiN on a chemically reactive substrate like gallium arsenide. Chemical bonding and physical phenomena like wetting and nucleation play a central role in controlling interfacial structure and thin film morphology.[1] These phenomena can often thwart attempts to fabricate particular microstructures. Basic surface science research is providing an understanding of the underlying mechanisms of thin film growth and structure on an atomic scale and this knowledge can be applied to overcome fundamental problems like the formation of unwanted phases by solid state reaction. It is well-known that many metals react on contact with semiconductors or undergo grain boundary-assisted diffusion at low temperatures when deposited in the high quality vacuums.[2] This is, in part, due to the chemical activity of metal atoms and also due to the fact that the incident beams of atoms, ions, molecules or clusters are often sufficiently energetic to cause the chemical reaction.[3] An obvious control strategy is the emplacement of an inert barrier layer to thermalize or neutralize the incoming species before they encounter the substrate. An implementation of this procedure is represented by the work of Waddill et al.[4] who used Xe films to control the morphology of metal-semiconductor interfaces and their electrical properties. The same group also found that the removal of oxygen from high T_c superconductors when metals are deposited can be minimized by cluster

PRECEDING PAGE BLANK NOT FILMED

deposition into a Xe buffer layer.[5] A condensed barrier layer can also be chemically active and serve as a reactive matrix for thin film growth. This new feature distinguishes this work from earlier efforts using condensed gas layers to modify thin film growth and, as we will discuss, provides some advantages that can be exploited in fabricating thin films.

The need for better methods of growing dielectric films on materials like gallium arsenide is motivated by the fact that compound semiconductors have comparatively poor native oxides.[6] Silicon is an ideal material for microelectronics, in part, because SiO_2, an excellent dielectric, can be grown by thermal oxidation. SiO_2 films are an essential element of every microelectronic device. For gallium arsenide, the obvious oxides Ga_2O_3 and As_2O_3 are thermodynamically unstable, difficult to grow and have relatively poor morphology and electrical properties when compared to SiO_2. Usable As_2O_3 layers can be produced in an As-rich atmosphere and the losses of As encountered during device fabrication can be compensated by going off stoichiometry. There is a need for growing better dielectric materials on compound semiconductors or simplifying the growth process.

THIN FILM PREPARATION

The details of the cryogenic reactive deposition procedure have been described elsewhere (refs. 7 and 8). The work was done in a stainless steel vacuum system designed for surface and interfacial science studies and the experimental procedure is contrained by the nature of this apparatus. However, the concept is applicable to any ultrahigh vacuum system designed for thin film deposition and may also be applicable to systems with lesser quality vacuums. The apparatus and procedure are shown schematically in Figure 1 . A semiconductor wafer was mounted on a molybdenum plate that was heated from behind by a 0 to 30 watt ceramic button heater. The molybdenum plate was attached to an OFHC copper cold finger that was cooled using a flowing He cryostat (APD Cryogenics Helitran) that can reach 15 K when cooled with helium or 78 K when cooled with liquid nitrogen.

Our work utilized a molecular beam epitaxy (MBE) grown GaAs(100) wafer approximately 1 cm x 1 cm which was cleaned and pre-etched using established methods or As capped. The As capped wafers were heated to 585 C to evaporate the capping layer. Ordered gallium arsenide surfaces require special sputter etching and annealing treatments. The specific procedures used normally produce gallium rich surfaces.[9] RHEED (reflection high energy electron diffraction) and photoemission data were used to establish the nature of the surface before the dielectric films were fabricated.

The substrates were heated and sputter etched at room temperature to prevent the cold trapping of impurities on the cold finger. During the cool-down to liquid nitrogen or helium temperatures, the button heater was used to keep the sample above 150 K to prevent condensation of significant amounts of residual gases like CO or water. When the cold finger and cryrostat were cold, the heater power was turned off to allow the waffer to cool down quickly. In the case of the reaction to form Al_2O_3, a jet of gaseous oxygen was directed on to the waffer using a needle doser while the temperature of the GaAs sample was held at 46-49 K. Aluminum vapor was then directed on to the sample from a resistively heated W basket at a rate of about 2 A/ min. The thickness of the Al was monitored using a quartz crystal microbalance. For the attempt to grow AlN, an ammonia film was first condensed and Al was then evaporated into it.

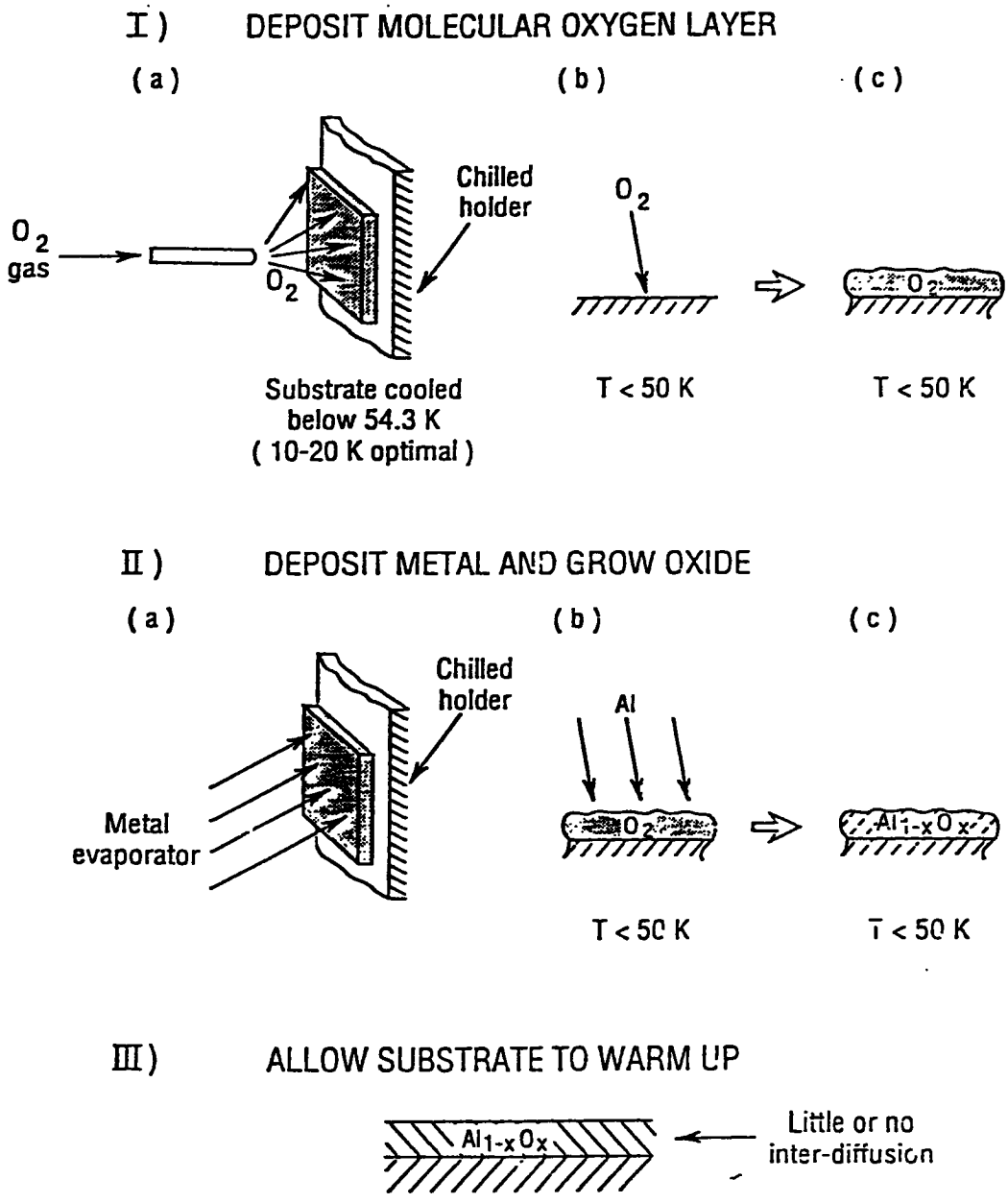


Figure 1. Schematic Diagram of the Thin Film Deposition Process

RESULTS AND DISCUSSION

Photoemission was used to characterize the thin metal compound layers formed on GaAs(100). The photoemission experiments have been reported elsewhere[7,8] but will be summarized here. In both cases to be illustrated, the Al inorganic compounds were formed on GaAs(100) below room temperature. It is well known that most metals react with an impurity free gallium arsenide surface and form metal arsenides or gallium alloy layers near the interface. Compound semiconductors are especially susceptable to solid state reaction because some of their surfaces contain both anions and cations.[10] Most metals (Al, Cu or Au) used as conductors react with gallium arsenide. Evidence of substrate disruption can be seen in core level photoemission data which is sensitive to the chemical state of the photo-ionized atom.[11] Specifically, the core level binding energy is related to the chemical state of the atom and, provided other factors are the same, the atoms in a number of different atomic environments will generate components that can be isolated in the spectra. Photoemission was used as a probe of thin film growth and composition because it is a more direct means of studying layers as thin a monolayer and is more sensitive than Rutherford backscattering.[12] Since the films are no where near the thickness required for device fabrication, the work described here is establishes the scientific feasability of the concept and further work is needed to make a technologically useful dielectric film.

For the growth of Al_2O_3 films on GaAs(100), the experiment consisted of three steps: (1) the condensing molecular O_2 on the surface, (2) vapor deposition of Al into the solid O_2 and (3) the heating of the sample to remove excess O_2.

In figures 2 and 3, we show valence band and core level data for the formation of aluminum oxide on GaAs(100) at 49 K. The results for the clean surface (Figure 2) are similar to those reported in many papers.[13] The valence band maximum (VBM) was found 0.42 eV below the Fermi level which indicates the surface is pinned. The photoelectron energies of the features shown in the experimental data are referenced to the VBM. Close examination of RHEED pattern, valence band photoemission spectra and the Ga 3d/As 3d ratios which were 0.48 indicates this surface is the gallium-rich (4x6) surface.

After the sample was cooled to 49 K, a jet of O_2 was directed on the sample and a thin O_2 film condensed on the sample surface. The triple point of O_2 is near 54 K and the vapor pressure at the critical point is 1.1 torr.[14] Being close to the critical point for O_2, limited the thickness of the O_2 layer to that of a weakly chemisorbed monolayer. Additional cooling would condense a much thicker O_2 layer. However, the thickness of O_2 was sufficent to make a thin of layer of aluminum oxide and very thin layers are frequently made by MBE. O_2 has a distinct valence electronic structure which is reflected in the photoemission spectra of the valence band. After the deposition of O_2, the valence band shows peaks identifed by Frankel et al.[15] with the $3\sigma_g$ (13.6, 11.2 eV), the $1\pi_u$ (9.8 eV) and the $1\pi_g$ (5.2 eV) molecular orbitals. The addition of Al vapor causes the reaction to form an aluminum oxide that can be identified on the basis of the Al 2p core level shift (Figure 3) as an amorphous Al_2O_3.[16] The molecular photoemission features disappear and are replaced by a broad peak about 6-8 eV binding energy which is also seen when metallic Al surfaces are oxidized.

There are a number of ways to make a sapphire film and the films produced have a variety of morphologies and compositions. A challenging problem with a material like gallium

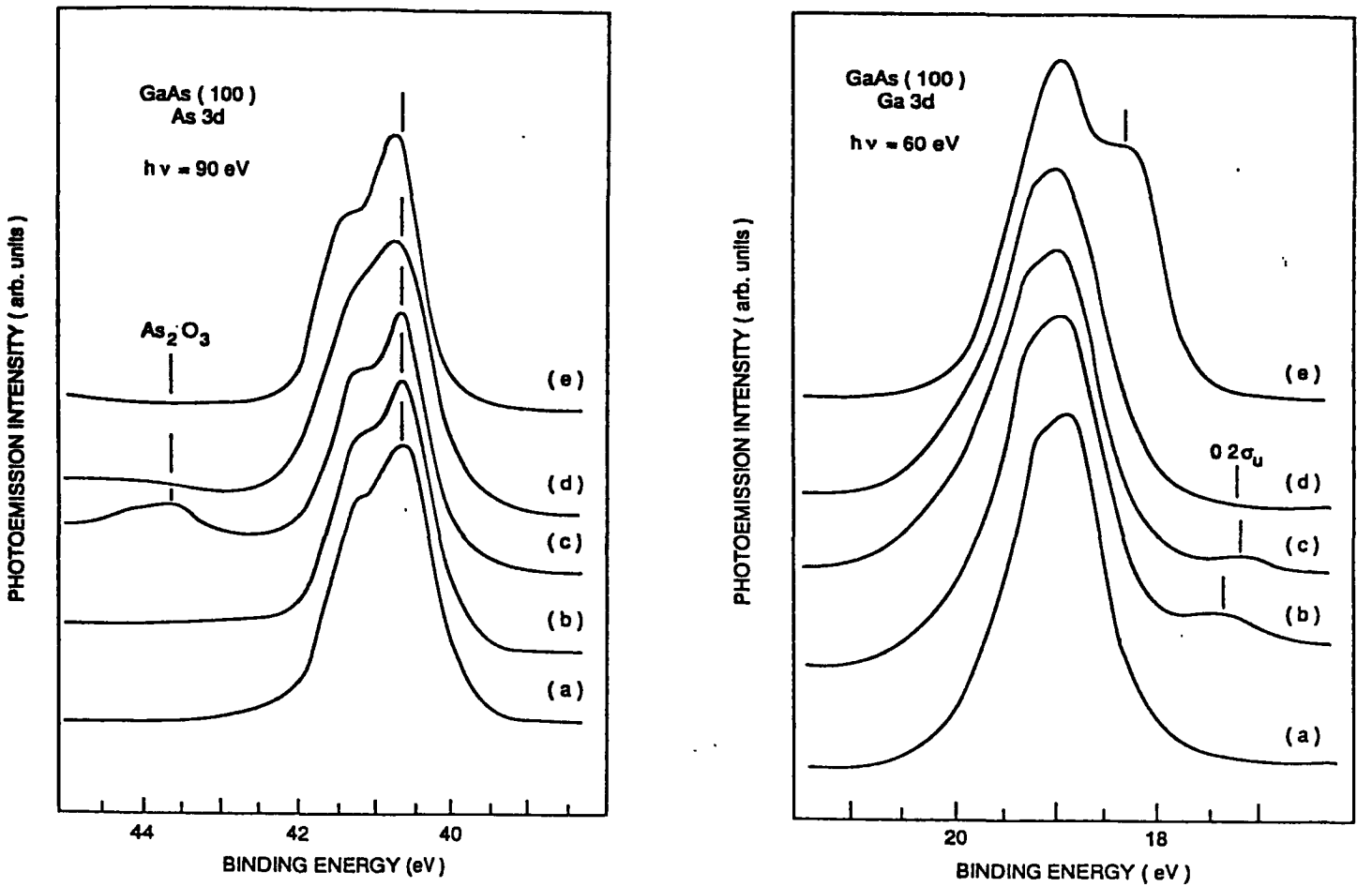


Figure 3. As 3d and Ga 3d core level photoemission spectra for the reactive deposition of Al into solid oxygen. For both figures (a) represents GaAs(100), (b) the same surface covered by O_2, (c) the same surface after the deposition of 2A of Al, (d) 10 A of Al A and (e) shows the effect of depositing 10 A of Al on GaAs(100) at 300 K.

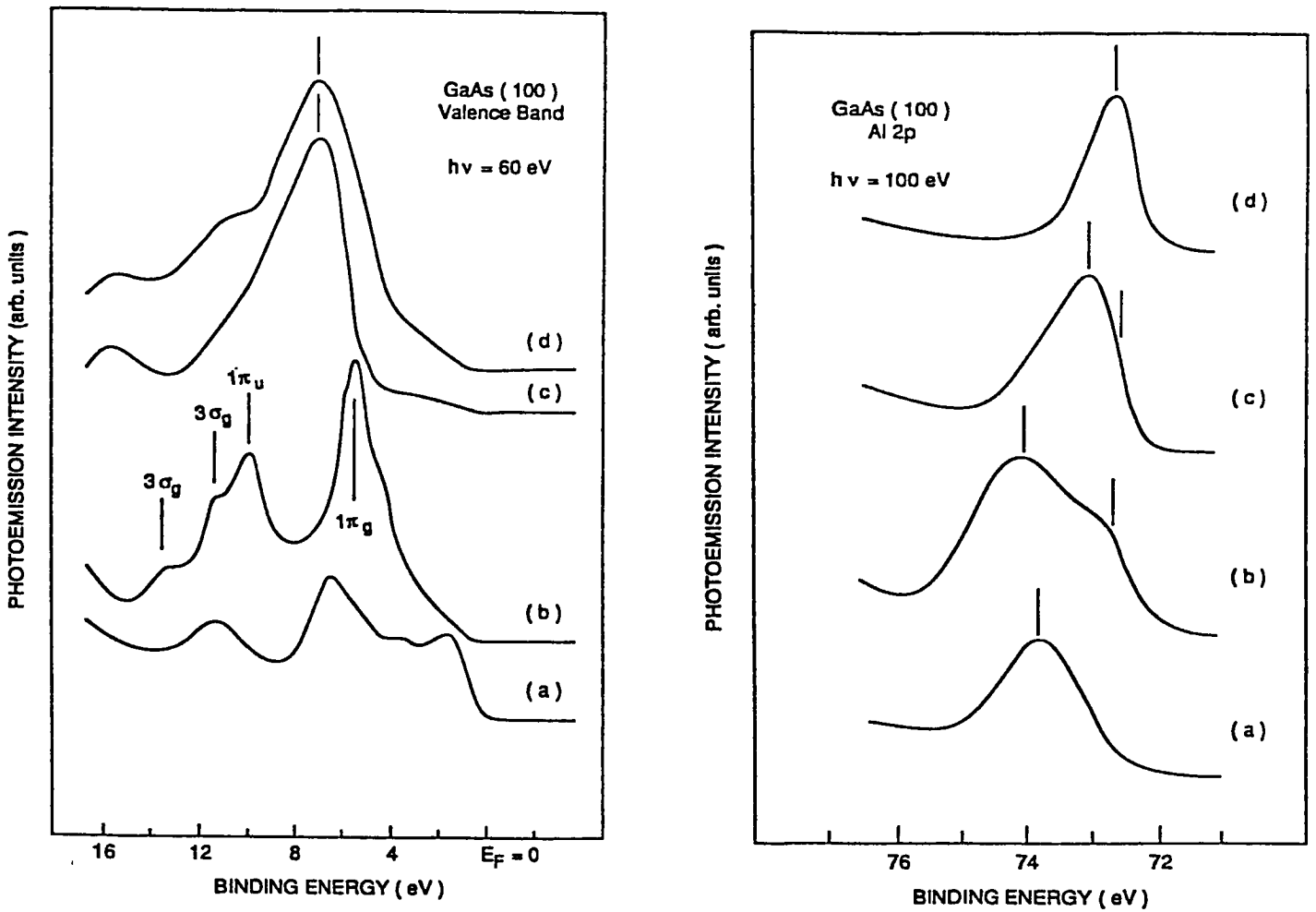


Figure 2. Valence and Al 2p core level photoemission spectra for the reactive deposition of Al into solid oxygen. For the valence band spectra (a) represents GaAs(100), (b) the same surface covered by O_2, (c) the same surface after the deposition of 2 A of Al , and (d) shows the surface upon warming to 300 K. For the Al 2p spectra (a) shows the Al oxide formed by depositing 1 A of Al, (b) 2 A of Al (c) the effect heating to 300 K and (d) the Al 2p spectrum for aluminum.

arsenide is the fact that the substate can react with the incoming species that make the thin film. Alternative methods of making Al_2O_3 films include oxidizing predeposited aluminum films, reactive evaporation, plasma-assisted chemical vapor deposition from metal organic precursors (PCVD), laser ablation of Al_2O_3 and radio frequency sputtering of Al_2O_3 target. All of the above techniques have drawbacks that can be largely eliminated by our new technique. For example, the plasmas generated by sputtering and PCVD heat and damage fragile substrates.

The Ga 3d and As 3d core level photoemission data in Figure 3 show nature of the dielectric/GaAs interface during the growth process, provide clues concerning the adhesion of the dielectric and indicate the extent of reaction between the dielectric and gallium arsenide. Extensive interdiffusion is seen when aluminum is deposited on the substrate and is evidenced in the spectra by the development of a second Ga 3d core level component (curve 3(e)) at lower binding energy that can be associated with an intermixed Al-Ga phase. An Al_2O_3 layer produced by the oxidation of a predeposited Al film would have a Ga oxide impurity that might decompose when heated. It is emphasised at this point that the aluminum oxide/GaAs(100) interface prepared by cryogenic reactive deposition lacks the Ga-Al phase seen when Al is directly deposited on GaAs. A small peak is seen about 3.1 eV above the As 3d peak. This component is due to a small amount of As_2O_3 that is produced by the reaction of arsenic with the atomic oxygen liberated during the reaction of O_2 with Al. This peak is reduced when the O_2 is used up and the Al reacts with As_2O_3 to form Al oxide and free As. We interpret the core level data as indicating that a relatively pure amorphous Al_2O_3 film with a sharp dielectric/GaAs(100) interface can be grown at temperatures of 50 K by the cryogenic reactive evaporation technique. The small scale modification of the gallium arsenide surface shows that there is bonding between the dielectric material and gallium arsenide. Aluminum oxide layers of useful thickness could be grown on a thin O_2 barrier film by co-deposition of aluminum and O_2 or by pulsed deposition of the two reactants separately.

The other dielectric deposited on GaAs(100) by the cryogenic reactive evaporation technique was aluminum nitride. CV measurements by Mizuta et al.[17] suggest that the interface density of states ($< 10^{12}$ /eVcm^2) is small between AlN and GaAs(100) making it a candidate for GaAs metal-insulator-semiconductor (MIS) system. AlN is also used as a capping material for gallium arsenide. AlN can be grown by a variety of techniques including but not limited to metal organic chemical vapor deposition (MOCVD), ion implantation of nitrogen into Al films and reactive molecular beam epitaxy. We attempted to grow AlN using cryogenically condensed ammonia.

Ammonia can be condensed on the surface at liquid nitrogen temperature (77 K) . The same procedures were used in this case as in the earlier experiment to make aluminum oxide. The chemistry of aluminum and ammonia is more complicated because a number of compounds can be formed between Al and ammonia or its fragments NH_2 and NH. Figure 4 shows valence band and Al 2p data for the reaction of Al with solid NH_3 on GaAs(100). Like O_2 , the condensation of NH_3 changes the valence band region. Peaks at -7.0 and -12.2 eV binding energy are assigned to the NH_3 $3a_1$ and 1e orbitals, respectively. We exposed the surface to 20 L of NH3 and based on the reduction in the intensity of the Ga and As 3d core levels estimate the thickness of the molecular film as being 3 monolayers.

The deposition of 10 A of Al causes the NH_3 features to shift about 0.8 eV to higher binding energy. Unlike O_2, the peaks are not disrupted and we conclude that NH_3 is retained

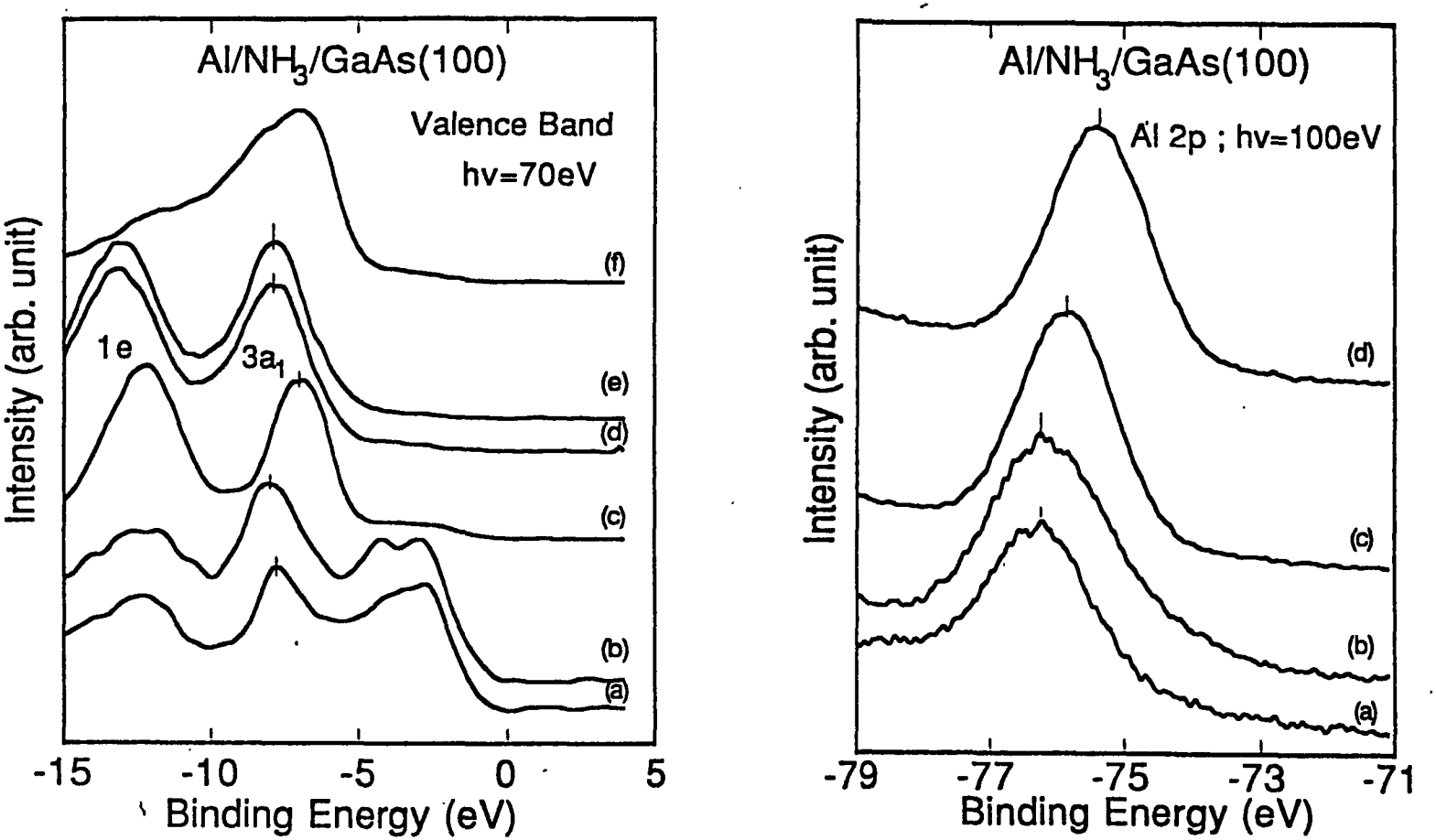


Figure 4. Valence and Al 2p core level photoemission spectra for the reactive deposition of Al into solid ammonia. For the valence band spectra (a and b) represent GaAs(100), (c) the same surface covered by NH_3, (d) the same surface after the deposition of 10 A of Al , (e) 20 A of Al and (f) shows the surface upon warming to 300 K. For the Al 2p spectra (a) shows the Al-NH_3 phase formed by depositing 10 A of Al, (b) 20 A of Al (c) the effect heating to 300 K and (d) 600 K (300 C).

as a bonded species in the Al-NH_3 compound. The deposition of a further 20 A of Al caused no changes in the valence band. Large scale changes are seen in the valence band when the sample is heated by allowing the cold finger to return to room temperature. The NH_3-related features disappear and a large broad feature peaking near -5.5 eV binding energy is seen. The shape of the valence band is the same as that obtained for AlN. There is a threshold temperature for AlN formation between 78 and 300 K. The large Al 2p core level (Figure 4) shift (2-3 eV) indicates that Al is chemically bonded to the NH_3. Thus a $Al(NH_3)_x$ precusor or aduct forms at low temperatures. Heating is required to decompose this to form the nitride, but as will be briefly discussed in the conclusion, processing using energetic photon or electron beams should also provide a method of generating this compound.

The As 3d and Ga 3d core levels show changes that reinforce the conclusions drawn from the valence level data and provide information about the AlN/GaAs(100) interface. The As 3d core level is broadened and the surface core level component is reduced when NH_3 is condensed. This can be connected with the chemisorption of the ammonia and the elimination of unsaturated (dangling) surface bonds. After 10 A of Al are deposited, the As 3d again broadens. This may be due to the formation of AsN or AlAs depending on whether hydrogen , an NH_x fragment or nitrogen is released during the insertion of Al into NH_3 or Al reaches a reaction site at the interface. The Ga 3d peak is also broadened but on the higher not lower binding energy side which is seen when Al is deposited and we conclude that AlAs has not formed at this stage of film growth. These changes are more pronounced after 20 A of Al are deposited. In addition, at the 20 A Al coverage, a small Ga 3d component associated with the Ga-Al phase begins to be seen. Heating the surface to 300 K, reverses some of the changes in the substrate seen when Al is inserted in to the NH_3 layer. The width of the As 3d decreases and the shape is comparable to that seen for a clean surface. The Ga 3d peak remains broadened and indicates that the GaN phase at the interface is stable at 300 K.

Both the Al oxide and AlN layers were heated to temperatures just below the threshold fc- GaAs substrate disruption and both films proved to be stable as evidenced by the lack of changes in the photoemission spectra.

SUMMARY AND CONCLUSIONS

From surface sensitive photoemission studies, we conclude that stable dielectric layers can be grown on a compound semiconductor by reactive deposition into cryogenically condensed molecular solids. In the specific examples, Al oxide and Al nitride were formed by reaction with O_2 at 49 K and NH_3 between 78 and 300 K. We believe the technique can be scaled up to make coatings of useful thicknesses and point out that the fabrication of these films under novel conditions provides an opportunity to make such films with morphologies that vary from those produced by more widely used techniques. Work has continued at Brookhaven using condensed molecular solids. Intense EUV and soft x-ray beams have been directed into condensed mixtures of gases like diborane and ammonia to produce coatings who spectroscopic properties resemble boron nitride.[18] Work is also proceeding using metal organics like tri-methyl aluminum to make compounds like AlN. We feel that condensed molecular films might play a role in advanced EUV and x-ray lithographic processes because molecular solids can be used to enhance etching and selective thin film deposition. The technique also has promise for depositng coatings like aluminum oxide on polymer surfaces for use as a protective layer. Another area that deserves closer examination is the use of this

process to deposit dielectric films on oxide surfaces like $YBa_2Cu_3O_{7-\delta}$ for use in planar tunneling junctions.

Acknowledgements

This work was supported by the United States Department of Energy Office of Basic Energy Sciences under Contract No. DE-AC02-76CH00016. Y. Gao also acknowledges the support of the Alfred P. Sloan Foundation.

REFERENCES

1. M. A. Herman and H. Sitter, Molecular Beam Epitaxy: Fundamentals and Current Status, (Springer-Verlag, Berlin, 1989) p 215-235.
2. N. Martenssen in Analytical Techniques for Thin Films, Treatise on Materials Science and Technology, Vol 27, K. N. Tu and R. Rosenberg eds., (Academic Press, Boston, 1988) p65-107.
3. G. D. Waddill, I. M. Vitomirov, C. M. Aldao, S. G. Anderson, C. Capasso, J. H. Weaver and Z. Liliental-Weber, Phys. Rev. B **41**, 5293 (1990).
4. G. D. Waddill, I. M. Vitomirov, C. M. Aldao and J. H. Weaver, Phys. Rev. Lett. **62**, 1568 (1989).
5. T. R. Ohno, Y-N. Yang, J. H. Weaver, Y. Kimachi and Y. Hidaka, Appl. Phys. Lett. **57**, 718 (1990).
6. S. K. Ghandi, VLSI Fabrication Principles: Silicon and Gallium Arsenide, (Wiley, New York, 1983) p. 416-475.
7. Y. Gao, C. P. Lusignan, M. W. Ruckman and M. Strongin, J. Appl. Phys. **67**, 7148 (1990).
8. K. T. Park, J. Cao, Y. Gao, G. W. Wicks and M. W. Ruckman, J. Appl. Phys. **70**, 2623 (1991).
9. T. Miller and T-C. Chiang, Phys. Rev. B **29**, 7034 (1984).
10. A. Kahn, Surf. Sci. Rep. **3**, 193 (1983).
11. D. Briggs in Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy, D. Briggs and M. P. Seah eds., (Wiley, New York, 1983) p. 359-392.
12. L. C. Feldman and J. W. Mayer Jr., Fundamentals of Surface and Thin Film Analysis, (North-Holland, New York, 1986) p. 13-28.
13. T-C. Chiang, R. Ludeke, M. Sono, G. Landgren, F. J. Himpsel and D. E. Eastman, Phys. Rev. B **27**, 4770 (1983).; R. Z. Bachrach, R. S. Bauer, P. Chianandia and G. V. Hanson, J. Vac. Sci. Technol. **19**, 335 (1981).
14. The Matheson Unabridged Gas Data Book, Matheson Gas Products, Inc. (1974).
15. D. J. Frankel, Y. Yunkun, R. Avici and G. J. Lapeyre, J. Vac. Sci. Technol., A **1**, 679 (1983).
16. D. E. Halverson and D. L. Cocke, J. Vac. Sci. Technol. , A **7**, 40 (1989); R. M. Henry, B. W. Walker and P. C. Stair, Solid State Commun. **42**, 43 (1982).
17. M. Mizuta, S. Fujienda, T. Jisukawa and Y. Matsumoto in Gallium Arsenide and Related Compounds, W. T. Lindley ed., (Institute of Physics, Bristol U. K. , 1987) p 153.
18. D. Strongin, M. W. Ruckman, J. Mowlem and M. Strongin (unpublished).

N92-22702

ADVANCED SILICON ON INSULATOR TECHNOLOGY

D. Godbey
Research Chemist
Naval Research Laboratory
Electronics Science and Technology Division
Code 6812
Washington, D.C. 20375-5000

H. Hughes
Research Physicist
Naval Research Laboratory
Electronics Science and Technology Division
Code 6812
Washington, D.C. 20375-5000

F. Kub
Research Engineer
Naval Research Laboratory
Electronics Science and Technology Division
Code 6812
Washington, D.C. 20375-5000

ABSTRACT

Undoped, thin-layer silicon-on-insulator has been fabricated using wafer bonding and selective etching techniques employing an MBE grown $Si_{0.7}Ge_{0.3}$ layer as an etch stop. Defect free, undoped 200-350 nm silicon layers over silicon dioxide are routinely fabricated using this procedure. A new selective silicon-germanium etch has been developed that significantly improves the ease of fabrication of the BESOI material.

INTRODUCTION

Silicon on insulator (SOI) technology is evolving as the older silicon on sapphire technology is replaced by newer technologies capable of supporting smaller geometry devices. Several technologies are competing to produce the next generation of SOI materials, and among the most important are silicon implanted with oxygen (SIMOX) and bond and etch back silicon-on-insulator (BESOI). There is also a subset of BESOI technologies under study that uses a combination of thinning techniques and different etch stops. The etch stops currently under study in the BESOI community include a heavily boron doped silicon layer [1] and a silicon-germanium alloy [2,3] as well as others.

Silicon on insulator has been identified as one of the candidate substrate materials for VLSI circuits in the decade beginning in the year 2000 [4]. Advantages of the SOI VLSI technology compared to conventional bulk VLSI technology include higher packing density, simpler process, and higher performance. Of the available SOI technologies including SIMOX and zone melt recrystallization, BESOI offers advantages in terms of lower defect density and the potential for lower manufacturing cost for large area (6, 8, or 12 inch) SOI wafers.

This work has focused on BESOI fabrication because the buried oxide, unlike with SIMOX material, can be hardened for operation in a radiation environment. As a result, large threshold voltage shifts are experienced in SIMOX devices following irradiation by ionizing sources. An added advantage is that the defect free BESOI device layers are better suited for high voltage and high temperature applications than SIMOX layers. BESOI utilizing a SiGe etch stop is preferred over BESOI employing boron or carbon doped silicon etch stops because of defects induced in the silicon device layer by the presence of these etch stops. An additional advantage that BESOI with the SiGe etch stop has over the boron etch stop process is the elimination of residual electrically active elements (boron) in the device layer following the etch stop removal.

These techniques are also compatible for the fabrication of thin silicon films such as silicon membranes, bolometers, and other devices and structures requiring free standing thin film silicon. Other potential applications include high voltage-high temperature power devices, backside illuminated thinned CCD imagers, and X-ray masks.

EXPERIMENTAL: BESOI Fabrication

A schematic of the BESOI fabrication process is shown in figure 1. Prime wafers were fabricated using molecular beam epitaxy (MBE) on Si(100) substrates that were cleaned chemically using a modified Shiraki procedure prior to being loaded into the MBE system. The oxide was removed *in vacuo* by heating to 800°C in a 0.1 Å/sec silicon flux. The structure was fabricated by growing the following layers: a 20 nm silicon buffer layer, a 60 nm $Si_{0.7}Ge_{0.3}$ alloy etch stop layer, and a 200-350 nm silicon epilayer. After MBE growth the wafers were evaluated by low energy electron diffraction (LEED) and optical microscopy with Nomarski.

The MBE grown prime wafer and the handle wafer were bonded together by the following procedure. Thermal oxides were grown on both the MBE prime wafer (850°C) and the handle wafer. The oxides were both hydrophilyzed in an NH_3:H_2O_2:H_2O solution, and then stacked together oxide to oxide in a microclean environment. Pre- bonding between the two wafers was established at room temperature [3]. The prebonded wafers were inspected using an infrared imaging system to determine the quality of the bond, and unqualified wafers (those showing voids) were reworked. Qualified wafers were annealed at 850°C, and this treatment was sufficient to produce a bond strength that allowed mechanical grinding and polishing of the wafer. The 850°C temperature was sufficiently high to cause some relaxation of the alloy etch-stop layer. As shown later, it did not have a negative impact on the stopping capability of the etch stop. In addition, dislocations formed by relaxation in the etch layer did not propagate into the epitaxial layer as shown by plan view transmission electron microscopy (TEM)[2]. Since no threading dislocations were observed by TEM, an estimate of the upper limit of dislocations in the Si epilayer is set at 10^4 cm^{-2}.

After bonding the prime and handle wafers, the backside Si of the prime wafer and the etch stop must be removed. The back side of the bonded prime wafer was thinned to the desired SOI thickness (200-350 nm) by a combination of mechanical thinning followed by selective etching [3]. Mechanical thinning was employed to bring the backside silicon plus etch stop layer thickness to less than 2 μm before chemical thinning. First, precision diamond grinding was used to reduce the thickness of the active wafer to about 25 μm. The depth of the damaged layer was about 3 μm following this step, and a micrograph of this surface is shown in figure 2. The wafers were then transferred to a precision polisher built by ARACOR to thin the backside layer to less than 2 μm. Figure 3 shows a very smooth layer, indicating that the damage layer was less than 100 nm thick. Optical interference fringe patterns were used to determine the thickness variation of the remaining semiconductor layer. For this sample, the film thickness was determined to vary between 1.5 and 2 μm. Further thinning used a selective silicon etch composed of 100 g. KOH, 4 g. $K_2Cr_2O_7$, 100 ml. propanol, and 400 ml. H_2O [2,3], which removed the silicon selectively from the etch stop layer. The $Si_{0.7}Ge_{0.3}$ alloy etch stop layer was selectively removed by a HNO_3:H_2O:HF(0.5%) solution, 35:20:10 vol:vol [5]. The selective etches for Si and SiGe are discussed below.

Following the removal of the etch stop layer, the finished silicon on insulator material is obtained. A cross sectional transmission electron micrograph of a 200 nm silicon on insulator film is presented in figure 4.

RESULTS AND DISCUSSION

The etches used for the BESOI process must have a high selectivity, defined as the ratio of the etch rate of the top layer to the etch rate of the sublayer. For example, the etch used to remove the remaining Si above the $Si_{0.7}Ge_{0.3}$ etch stop layer must have a fast etch rate for Si and a slow rate for $Si_{0.7}Ge_{0.3}$. Conversely, the etch which is used to remove the etch stop, must have a fast etch rate for $Si_{0.7}Ge_{0.3}$ and a slow rate for Si. An additional constraint on the etches is that the final surface must be left polished, suitable for device processing.

The selective silicon etch showed the behavior indicated in figure 5. The slope of the steep curves on the left side of the plot is the etch rate of the silicon layer. The rate obtained was 19 nm/minute and did not change following heat treatment to 850°C as shown in the plot. The shallow curves on the right give the etch rates

through the alloy etch stop layer and were 0.8 and 1.1 nm/minute for the as grown alloy layer and the layer heated to 850°C for 30 minutes respectively. The selectivity was approximately 20:1.

The selective removal of the $Si_{0.7}Ge_{0.3}$ alloy etch stop layer is shown in figure 6. The slope of the left side of the curve gives the etch rate through the etch stop layer which was 41 nm/minute. The etch rate through the underlying silicon layer was 0.4 nm/minute, giving a selectivity for the etch stop removal of 100:1. The alloy etch left the surface polished, as required.

Spreading resistance profiling was done on a 200 nm Si BESOI layer. The results showed the fabricated device layers to have a residual doping level of $8x10^{14}$ cm^{-3} p-type. Thus we have the basis for a fully depleted BESOI technology. MOS devices have been fabricated, and their performance is shown in figures 7 and 8 for n-channel and p-channel MOSFET's, respectively. The n-channel devices show the well known "kink" effect typical of thin film MOS devices. These devices were well behaved as shown, and future work will include research on the radiation tolerance of the NRL BESOI compared to other SOI technologies.

CONCLUSIONS

The use of an epitaxial $Si_{0.7}Ge_{0.3}$ layer etch stop in the fabrication of BESOI has been developed. This technique utilizes thermal and deposited oxides to form the buried oxide layer, thereby enabling the use of standard radiation hardening techniques in the growth of the buried oxide. The use of an epitaxial $Si_{0.7}Ge_{0.3}$ layer as an etch stop results in a defect free and undoped silicon on insulator film. The silicon film can be grown to any thickness desired, and silicon films in the 200-350 nm range are routinely fabricated.

ACKNOWLEDGMENT

This work was partially carried out under NRL contract with ARACOR of Sunnyvale, CA. This work was funded by the Defense Nuclear Agency.

REFERENCES

[1] W.P. Maszara and G. Goetz and A. Caviglia and J.B., McKitterick, J. Appl. Phys., 64, 4943, 1988.
[2] D. Godbey, H. Hughes, F. Kub, M. Twigg, L. Palkuti, P. Leonov, J. Wang, Appl. Phys. Lett., 56, 373, 1990.
[3] D. Godbey, M. Twigg, H. Hughes, F. Kub, L. Palkuti, P. Leonov, J. Wang, J. Electrochem. Soc., 137, 3219, 1990.
[4] National Advisory Council, Microtech 2000 Report, August 1991, Research Triangle, N. Carolina.
[5] A. Krist, D. Godbey, N. Green, Appl. Phys. Lett., 58, 1899, 1991.

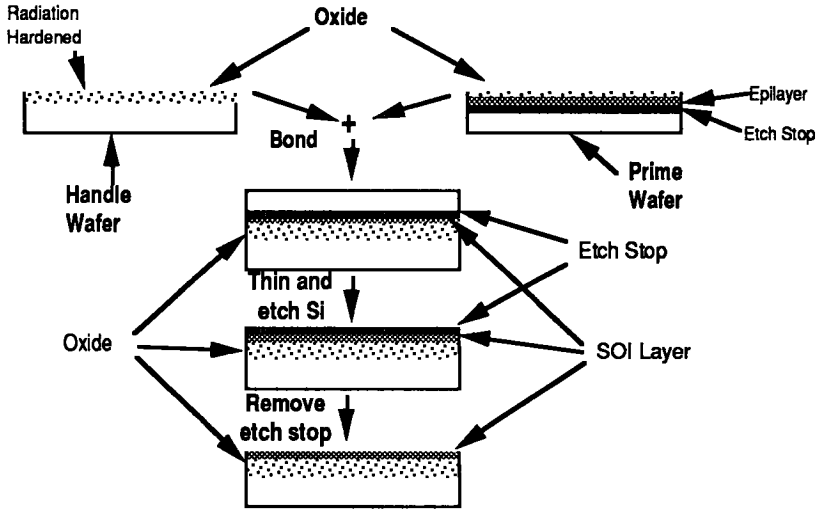


Figure 1. Schematic representation of bond and etch back silicon on insulator fabrication.

Figure 2. Active side of bonded wafer after being ground to thickness of 25 μm showing the 3 μm surface damage.

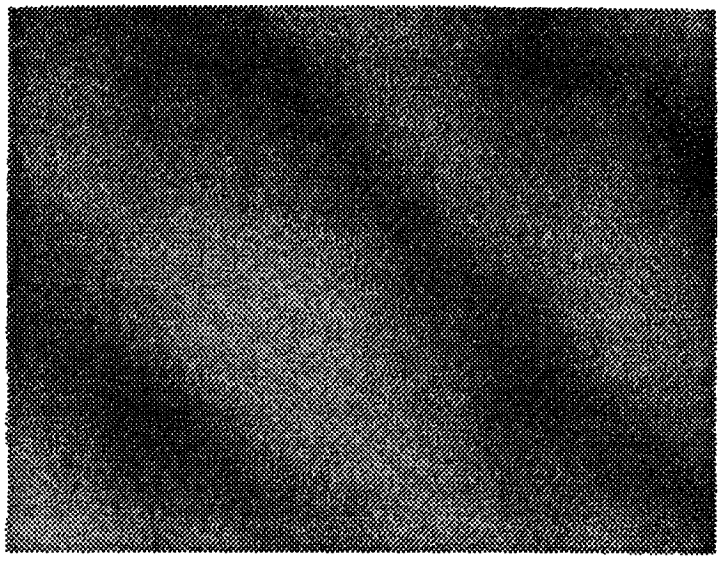


Figure 3. Active side of bonded wafer after being precision polished to a thickness of 2-1.5 μm showing a smooth surface and interference fringes.

Figure 4. Silicon on insulator fabricated by bond and etch back using a SiGe alloy as etch stop showing a 200 nm SOI layer on a 920 nm oxide layer.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

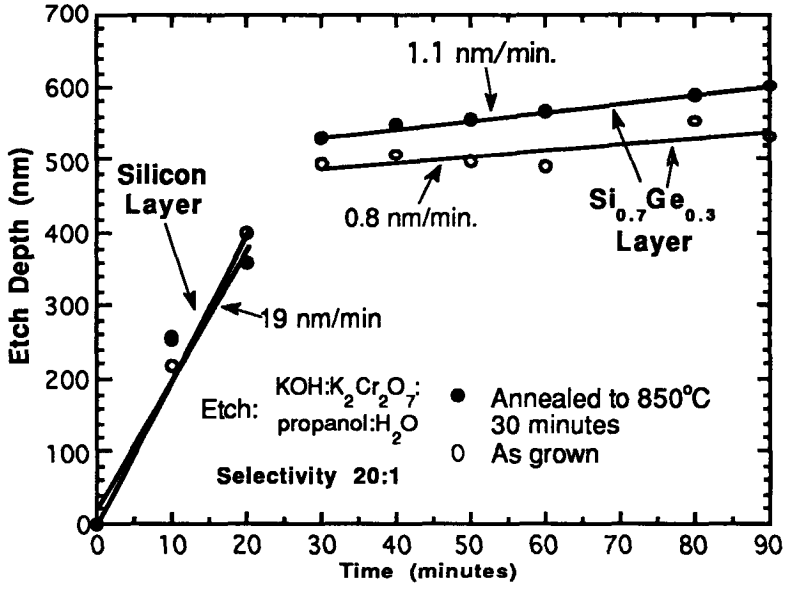


Figure 5. Silicon selective etch. The steep slope gives the silicon etch rate, while the shallow slope gives the alloy etch rate.

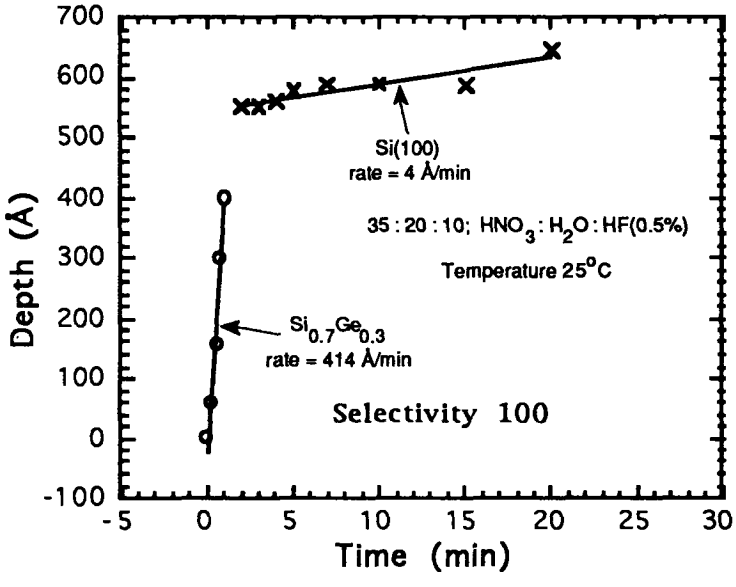


Figure 6. SiGe (30% Ge) selective etch. The steep slope gives the alloy etch rate, while the shallow slope gives the silicon etch rate.

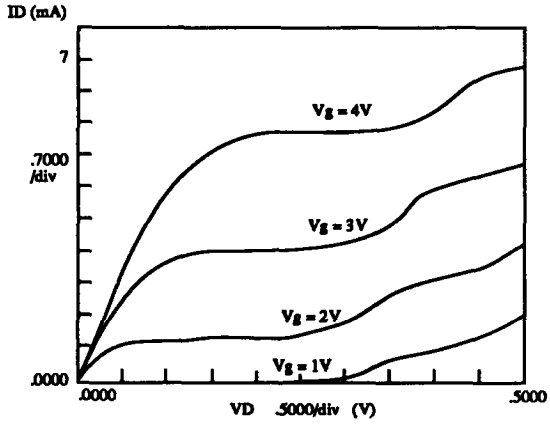


Figure 7. Drain characteristics of the 1.0 μm n-channel MOSFET. The floating-body related kink effect is seen for this device.

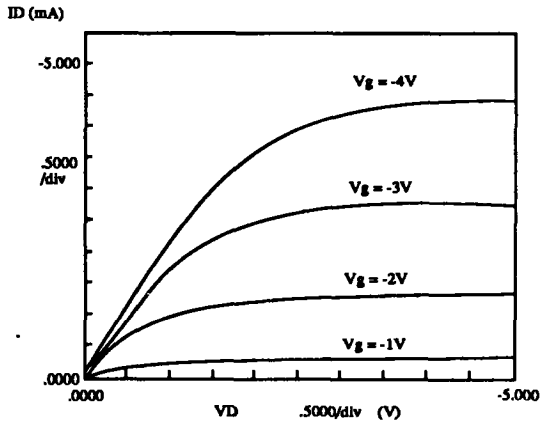


Figure 8. Drain characteristics of the 1.0 μm p-channel MOSFET. The applied bias to the substrate was +5 V during measurement to control the back-channel threshold.

N92-22703

PERFORMANCE OF A HIGH T_c SUPERCONDUCTING ULTRA-LOW LOSS MICROWAVE STRIPLINE FILTER

J. J. Bautista and G. Ortiz
Jet Propulsion Laboratory
Pasadena, CA 91109

C. Zahopoulos and S. Sridhar
Department of Physics, Northeastern University
Boston, MA 02115

M. Lanagan
Argonne National Laboratory
Argonne, IL 60439

ABSTRACT

This article reports successful fabrication of a five-pole interdigital stripline filter made of the 93-K superconductor ($Y_1Ba_2Cu_3O_y$) coated on a silver substrate, with center frequency of 8.5 GHz and an extremely high rejection ratio of 80 dB. The lowest insertion loss measured was 0.1 dB at 12 K, with a return loss of better than 16 dB, representing a significant improvement over a similar copper filter, and is comparable to low critical-temperature filters. The insertion loss appears to be limited by extrinsic factors, such as tuning mismatch and joint losses, and not by the superconducting material losses.

INTRODUCTION

Recent results of the intrinsic microwave properties of high-quality single crystals [1] and thin films [2] of the high critical-temperature (T_c) superconductors raise tremendous prospects for the applications of these materials in microwave devices. These results of the surface resistance imply significantly improved performance as compared with devices using conventional materials, such as copper. It is clear, however, that significant problems need to be solved in order to translate the results on small scale samples into realistic structures necessary for actual devices.

A bandpass filter is of great utility in systems limited in performance by radio frequency interference (RFI) at the input. This is particularly important in space communications, where the front-end amplifier is a delicate high-electron mobility transistor (HEMT) or maser amplifier with a very low noise temperature. The effect of incident RFI is primarily determined by the level and frequency of the RFI. Both in-band and out-of-band RFI can result in significant gain compression, noise temperature increase, and, in the case of the maser, spurious output signals. For example, a 0.1-dB increase in insertion loss can result in a 0.4-K increase in noise temperature, which can be serious in low noise systems. Therefore, a usable (out-of-band) RFI filter should have stringent specifications: a narrow bandpass, a high out-of-band rejection ratio (at least 80~dB), and an extremely low insertion loss (~0.1 dB) to avoid in-band signal attenuation and added noise.

FILTER STRUCTURE

This article describes the fabrication and performance of an 8.5-GHz bandpass filter made of silver and coated with $Y_1Ba_2Cu_3O_{7-x}$ (YBCO). An interdigital tunable stripline resonator structure was selected over a microstrip filter as the optimal design due to its favorable performance characteristics [3], and its potential ability to meet the design criteria specified above. This structure is compact and, with the exception of waveguide filters, it has the highest unloaded resonator (Q) among the commonly used structure [4]. The filter, shown in Fig. 1, consists of five transverse electromagnetic (TEM) mode stripline resonators. Each resonator is approximately one-quarter wavelength long at the midband frequency and is short-circuited at one end and open-circuited at the other. These resonators are placed between two ground plates which are attached to the filter

body by as many screws as possible in order to reduce losses at the joints. Bandpass tuning is accomplished by varying the capacitance of the resonators with the five adjustable screws opposite each of the five fingers. This particular filter was designed with the aid of the low-pass prototype synthesis methods outlined in [4]. It is a 0.05-dB equal-ripple bandpass filter with an equal-ripple bandwidth of 0.15 GHz centered at 8.5 GHz.

FABRICATION METHOD

Three different fabrication methods were considered: (1) making the entire filter out of bulk T_c superconductor; (2) coating a silver-plated copper filter with a thick YBCO film; and (3) coating a pure silver filter with a thick superconducting film. Despite experience with bulk high T_c structures [5], the first method, after some initial trials, was abandoned due to the complexity of the structure (many sharp edges, screw holes, threads, etc.). Some of the edges tended to chip, and the drilling of so many holes weakened or even destroyed the structure. In the second approach a copper filter was machined and silver-plated with a 500-micron thick film. This filter was subsequently coated with a thick YBCO film. The resultant film looked dark gray and was not superconducting. This process was repeated with a number of test pieces without the desired success. A possible explanation of this may be that the silver buffer layer degraded at high temperatures and the copper substrate reduced the thick YBCO film. The same result was obtained even when the sintering temperature was reduced from 920 deg C to 900 deg C. Finally, the third method attempted was very successful. It involved machining the filter out of 99.9 percent pure silver, and subsequently coating the filter with a thick YBCO film, as in the previous method. The resulting films looked black and exhibited sharp superconducting transitions with $T_c \sim$ \$92 K. By using this last method, three different filters were fabricated and tested, making appropriate improvements each time.

The YBCO compound was prepared via a solid state reaction by using yttrium oxide, barium carbonate, and copper oxide. Stochiometric amounts of the constituent materials were mixed and ballmilled in methanol for 16 hr. The slurry was dried and vacuum calcined [6] at 800 deg C for 4 hr in an oxygen pressure of 2.7×10^2 Pa. Thick films were fabricated by mixing the YBCO powder with an organic solvent, and a dispersant was added to improve the rheological properties. The suspension was applied to the silver filter substrate and dried at 80~deg~C for about 2~hr. The film was then sintered at 920 deg C for 4 hr in an oxygen partial pressure (PO_2) of 1.1×10^4 Pa and annealed at 450 deg C for 16 hr in 1 atmosphere of oxygen. Sintering in low PO_2 enhances [7] the sintering kinetics of YBCO. In addition, the melting point of silver is slightly higher in reduced PO_2. The resulting coating thickness was on the order of 50 microns.

MEASUREMENT RESULTS

An 8510B Hewlett Packard Network Analyzer with an S-parameter test set was used to precisely tune the filter and perform both the return and insertion loss measurements. The filter was originally tuned at room temperature to better than 20 dB of return loss across the frequency band of interest and subsequently cooled down to 12 K by using a closed-cycle refrigerator (CCR). Its temperature was monitored by two separate sensors attached to its body and the temperature-dependent data were taken as the filter was allowed to warm up slowly. The insertion loss of the coaxial lines inside the CCR are substrated from the data presented. These lines were separately characterized as a function of temperature for the frequency band of interest.

For an equal-ripple Chebyshev filter, the insertion loss (L_s) at midband is given by the expression:

$$L_s \text{ (dB)} = 8.686 \, (C_n/WQ_u) \qquad (1)$$

where Q_u is the unloaded resonator Q, W is the fractional bandwidth, and C_n is a coefficient determined by the filter order and its band ripple [2].

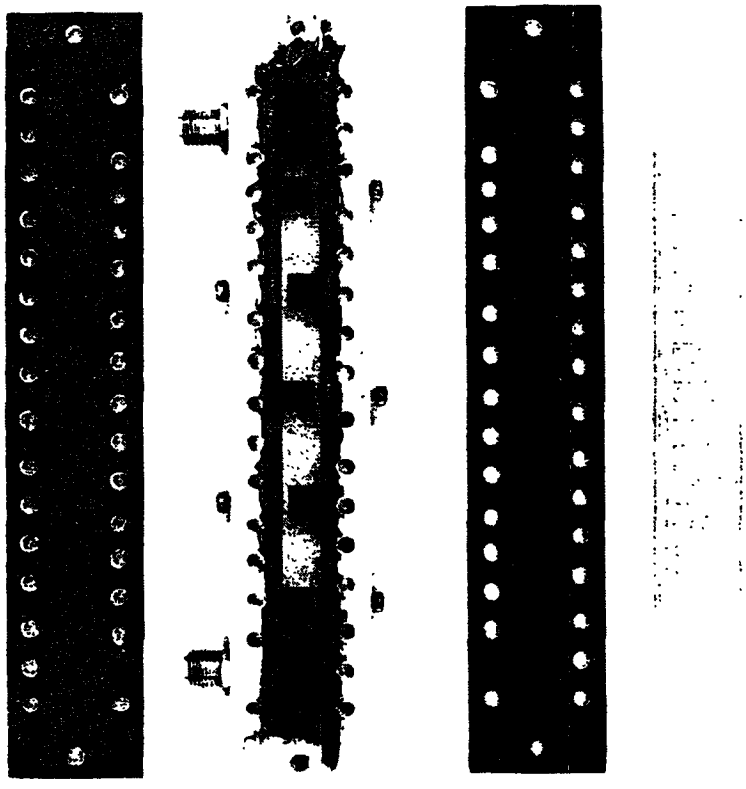


Fig. 1. The superconducting YBCO/Ag filter.

From Eq. (1), since Q_u \proportional to $1/R_s$, it is readily seen that the insertion loss of a filter is proportional to the surface resistance of the material it is made of. For conventional superconductors like Pb (T_c = 7.2 K) and NbTi (T_c = 9.8 K) at 4.2 K and at X-band frequencies, the surface resistance is known to be as much as three orders of magnitude lower than that of copper. The expected insertion loss for an ideal filter like the one considered here is therefore on the order of 10^{-4} dB. From the results on the surface resistance of polycrystalline and single crystal YBCO materials [1], the authors expect insertion losses on the order of 0.25 dB for the polycrystalline and no more than $10^{-3} dB for an ideal single-crystal filter.

Figure 2 shows the temperature dependence of the insertion loss of the three different silver and YBCO filters as a function of temperature. All the data were taken at an input continuous wave (cw) power level of 38 micro W. Qualitatively, all the curves look similar. A sharp transition at 93 K is observed as the film becomes superconducting. The insertion loss then tails off at about 70 K, at which point it starts decreasing linearly to 12 K. The lowest value reached is 0.69 dB, and if extrapolated down to 4.2 K, the insertion loss would be 0.61 dB, as compared with 0.55 dB for a similar copper filter. At lower values of input power, the filter exhibits losses lower than that of copper, as discussed later.

The data of Fig. 2 for the three filters that were constructed reveal an important feature, namely that the differences in performance among the three trials were not due to material properties (i.e., R_s), but rather due to nonoptimization of the final devices. This is evident if one superposes the three curves by subtracting constant (temperature-independent) offsets, whence the three curves become identical. Thus, in practice, the insertion loss is represented by $L_m(T) = L_s(T) + L_0$, where L_0 is temperature-independent and arises from connector mismatch, tuning, etc. When this was realized, it was possible to achieve the best results with filter 3 by improving the ground-plane contacts, and by carefully assembling and tuning the filter at room temperature to the lowest insertion loss achievable.

Earlier work [3] with NbTi (T_c = 9.8 K) filters substantiates the above conclusions for the high T_c filter. In the NbTi filter, it was discovered that poor ground-plane contacts can contribute as much as 0.5 dB to the insertion loss at cryogenic temperatures and could be minimized by using knife edges at the joints. The ultimate residual loss achieved with the NbTi filter was 0.10 dB, and it was concluded that this represented the loss due to the connectors.

Thermal cycling strongly affects the L_s data. For filter 3, L_s was found to increase after the first thermal cycle and following the filter assembly. After the first thermal cycle, L_s increased by 0.7 dB at room temperature, and the subsequent cool-down data showed an increase by the same amount. Following the final cryogenic measurement, it was found that room temperature disassembly and reassembly of the filter increased its loss by 3.0 dB. Thus, for future designs, care must be taken to reduce warping and dimensional changes in the structure due to the high temperatures involved in the fabrication process. In addition, engineering design changes will have to be incorporated to eliminate the contact losses at the ground planes.

The surface resistance of samples prepared in the exact same way as the filter was also measured as a function of temperature down to 77 K by using a 14-GHz microwave cavity, with the sample disk as an end plate. The R_s data showed the same qualitative temperature dependence that was observed for the insertion loss measurements.

As mentioned earlier, the insertion loss L_s was found to exhibit a strong input power dependence, even at very low input powers. Figure 3 shows L_s as a function of input power which was varied from 38 micron W to 40 pW. For the first cool down and very low input cw power levels (>1 nW) the filter loss approached very low values of about 0.05 dB. As indicated in Fig. 3, the maximum error at the low power levels and for very low values of L_s is about 0.3 dB, which represents a very conservative upper bound on the measurement.

Such strong dependence on the input power is somewhat puzzling, since the incident magnetic fields on the superconducting film are too weak to account for such an effect. A thermal gradient could, however, exist in the YBCO material at the ground-plane contacts. Although the silver surface was machined flat to within 0.001 in., irregularities caused by firing and uneven film thickness prevented perfect mechanical contact at the interface, and hence led to poor electrical and thermal performance. It is likely that power levels above 38 nW could be sufficient to locally heat the YBCO material at the shorted base of the resonators a few kelvin above the rest of the material. It should be recalled that the radio frequency current densities are highest at the shorted base of the resonators. This is consistent also with the observed thermal cycling degradation, since the stainless steel screws would have a slightly different coefficient of expansion than the silver body, leading to weaker contact force and hence higher losses.

Note several features of the device reported here that bear on comparisons with thin-film microstrip filters. As part of its design, the waveguide should possess the lowest losses achievable, in comparison with microstrip filter structures, which have higher losses. The five element configuration also provides a high out-of-band rejection (~80 dB), which is difficult to achieve with thin-film microstrip filters. Connectorization should, in principle, be easier here because of the metallic substrate. It was also noted that the very low insertion loss of the filter has forced careful evaluation of the procedures for measuring L_s, and indeed it is clear that even better performance will require more exact procedures. It should also be noted that at least in space communications, signal power levels are usually very low (~ nanowatts), which is the level at which this filter exhibits its lowest insertion loss (see Fig. 3).

The authors have also tested the filter in an actual system configuration designed to measure system noise temperature with a HEMT front-end amplifier. Details of the measurement will be presented elsewhere. At a physical system temperature of T_m = 20 K, and with the amplifier noise temperature of T_L = 20 K, the experiments yielded a noise temperature contribution of 4.6 K, which agrees well with that inferred from the measured insertion loss data.

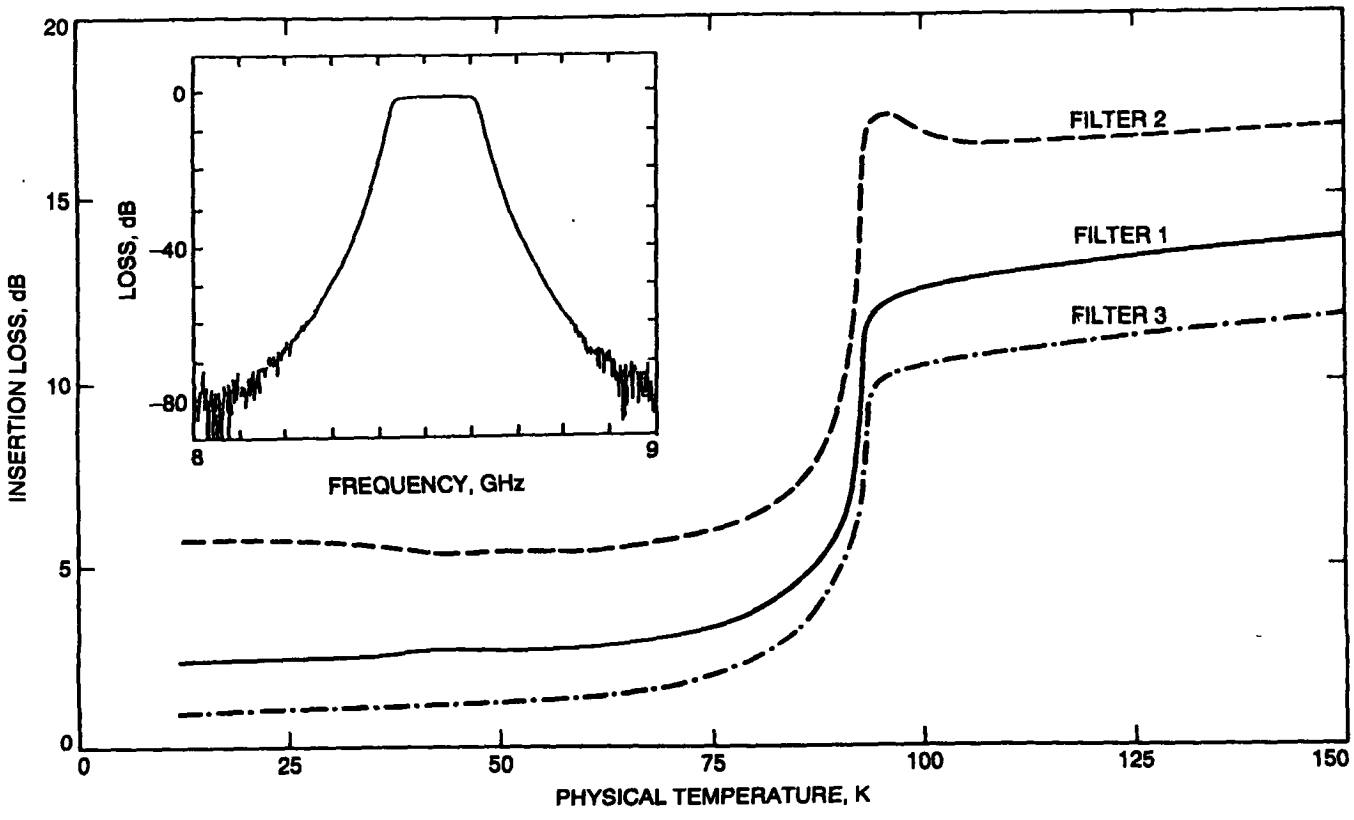


Fig. 2. Insertion loss (Ls) versus physical temperature (T) for three separate filters. The measurements were taken at an input power of 38 μW.

In conclusion, a stripline high T_c superconducting microwave bandpass filter wassuccessfully fabricated and tested. The lowest insertion loss measured was 0.05 dB, with an input return loss of better than 16 dB across the passband. The filter provides significant improvements over a comparable copper filter and is at present limited not by the superconducting material, but rather by design limitations possibly originating at the time of fabrication. Even at present, the filter is comparable to low T_c superconducting devices (with a distinct advantage over the latter in that operation is possible at elevated temperatures which are cost effective), and meets design criteria for the very low-noise communication systems of deep space applications.

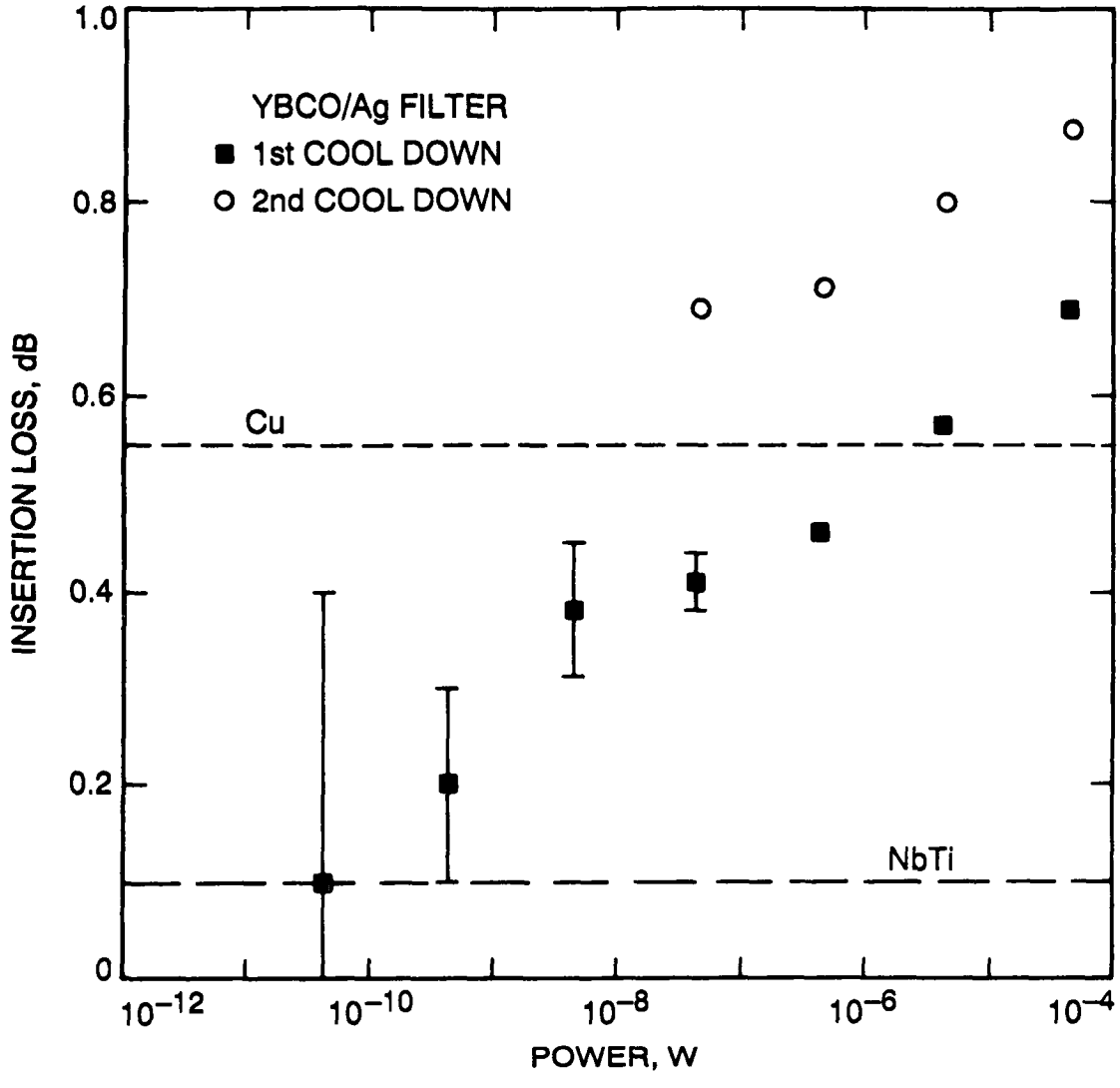


Fig. 3. **Insertion loss measurement at various power levels at 12 K for the YBCO/Ag filter, and similar Cu and NbTi filters.**

REFERENCES

[1] D.H. Wu, L. L. Kennedy, C. Zahopoulos, and S. Sridhar, "Characteristics and Growth of Single Crystals of $Y_1Ba_2Cu_3O_7$ with Superior Microwave Properties," *Appl. Phys. Letters*, vol. 55, no. 7, pp. 696–698, August 14, 1989.

[2] A. Inam, et al., "Microwave Properties of Highly Oriented $YBa_2Cu_3O_{7-x}$ Thin Films," *Appl. Phys. Letters*, vol. 56, no. 12, pp. 1178–1180, March 19, 1990.

[3] J. J. Bautista and S. M. Petty, "Superconducting NbTi and Pb (Cu) Bandpass Filters," *IEEE Trans. on Magnetics, MAG-21*, no. 2, pp. 640–643, March 1985.

[4] G. L.~Matthaei, L. Yound, and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Dedham, Massachusetts: Artech House Books, pp. 421–427, 1980.

[5] C. Zahopoulos, W. L. Kennedy, and S. Sridhar, "Performance of a Fully Superconducting Microwave Cavity Made of the High T_c Superconductor $Y_1Ba_2Cu_3O_x$," *Appl. Phys. Letters*, vol. 52, no. 25, pp. 2168, June 20, 1988.

[6] U. Balachandran, R. B. Poeppel, J. E. Emerson, S. A. Johnson, M. T. Lanagan, C. A. Youngdahl, D. Shi, and K. C. Goretta, "Synthesis of Phase-Pure Ortharhombic $YBa_2Cu_3O_x$ Under Two Oxygen Pressure," *Materials Letters*, vol. 8, no. 11, 12, pp. 454–456, November 1989.

[7] N. Chen, D. Shi, and K. C. Goretta, "Influence of Oxygen Concentration on Processing $YBa_2Cu_3O_{7-x}$," *J. Appl. Phys.*, vol. 66, no. 6, pp. 2485–2488, September 5, 1989.

N92-22704

An Adjustable RF Tuning Element for Microwave, Millimeter wave, and Submillimeter Wave Integrated Circuits

Victor M. Lubecke,[1] William R. McGrath,[2] and David B. Rutledge[1]

1. Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125
2. Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

ABSTRACT

Planar RF circuits are used in a wide range of applications from 1 GHz to 300 GHz including radar, communications, commercial RF test instruments, and remote sensing radiometers. These circuits, however, provide only fixed tuning elements. This lack of adjustability puts severe demands on circuit design procedures and materials parameters. We have developed a novel tuning element which can be incorporated into the design of a planar circuit in order to allow active, post-fabrication tuning by varying the electrical length of a coplanar strip transmission line. It consists of a series of thin plates which can slide in unison along the transmission line, and the size and spacing of the plates are designed to provide a large reflection of RF power over a useful frequency bandwidth. Tests of this structure at 1 GHz to 3 GHz showed that it produced a reflection coefficient greater than 0.90 over a 20% bandwidth. A 2 GHz circuit incorporating this tuning element was also tested to demonstrate practical tuning ranges. This structure can be fabricated for frequencies as high as 1000 GHz using existing micromachining techniques. Many commercial applications can profit from this micromechanical RF tuning element, as it will aid in extending microwave integrated circuit technology into the high millimeter wave and submillimeter wave bands by easing constraints on circuit technology.

INTRODUCTION

The vast field of RF electronics has long capitalized on the advantages of simple and cost effective planar circuit technology. From 1 GHz to 40 GHz, microstrip, coplanar strip, and slotted transmission lines are commonly used for signal distribution in hybrid and monolithic circuits used in radar, communications, and remote sensing applications. Typically, the transmission lines in these circuits are designed with a particular characteristic impedance, and some form of impedance transformation or tuning circuit is used to insure that the circuit components transfer the RF signal along these lines in the most efficient way. Ideally, knowledge of the operating impedance of each component allows for a fixed circuit design which incorporates this impedance transformation. In practice, however, it is quite difficult to accurately characterize these components, particularly at high frequencies. As a result, some post fabrication tuning in the form of circuit modification is required. This type of tuning is not easy and it is required more often as the design frequency is increased. It is therefore desirable to incorporate impedance matching elements in these circuits which can be readily fined tuned after fabrication in order to relax the constraint of component characterization, and thus allow the circuit design to be extended to higher frequencies.

In waveguide circuits, this tuning is accomplished with a mechanically adjustable backshort which is inserted into the waveguide. This provides a tuning stub with an adjustable electrical length. Waveguide, however, is often difficult to interface with planar components, and the dimensional requirements can make the waveguide circuit quite difficult and costly to fabricate. Analogously, an approach for a movable, noncontacting "planar backshort" which can be used to vary the electrical length of a coplanar strip transmission line tuning stub has been developed. This sliding circuit element allows for the design of an RF circuit which can be actively tuned after fabrication to achieve optimal performance.

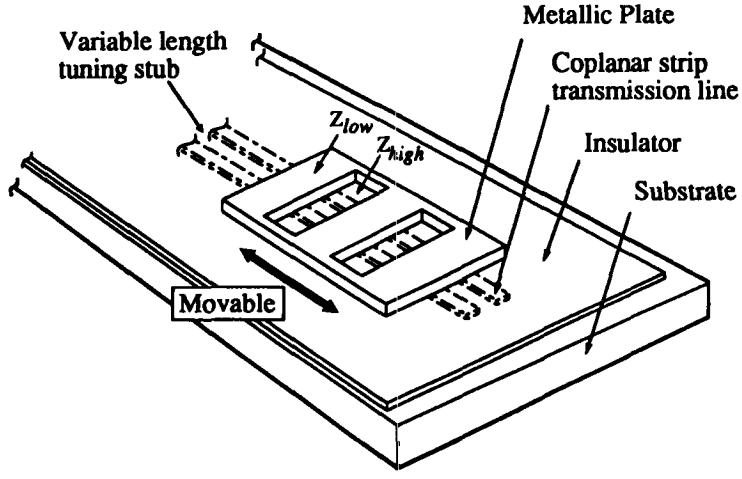


Fig. 1. Design of sliding backshort on coplanar strip transmission line. The holes in the metal plate create a series of successive high and low impedance sections that produce a large reflection of RF power.

THE PLANAR BACKSHORT

Placing a solid metallic plate across a coplanar strip transmission line, with a thin dielectric insulating layer in between, results in a reflection of RF power. Unfortunately, this reflection is not sufficiently large over a useful bandwidth for the plate to be used as a backshort in the design of a practical tuning circuit. This "sandwich" configuration does, however, result in a section of lower impedance transmission line. Quarter-wavelength sections of this line can be cascaded alternately with uncovered sections of "high" impedance line to create a series of impedance transformations which ultimately result in a very low impedance. When used as a backshort on a transmission line, this cascade produces an even larger reflection of RF power than the solid plate. We constructed and optimized such a planar backshort and it's geometry is illustrated in Fig. 1. It consists of a thin metal plate with rectangular holes of the proper dimension and spacing. A thin insulator keeps the plate from contacting the transmission line which reduces wear and allows the backshort to slide freely when pushed.

MODEL MEASUREMENTS

We built a large scale model of the sliding backshort and measured the magnitude of the reflection coefficient with an HP 8510B network analyzer over a frequency range of 1 GHz to 5 GHz. This measurement required a transition between the coaxial input of the network analyzer and the coplanar line. Unfortunately, no standard transition with low VSWR is readily available. For this reason, we devised several distinct measurement techniques in order to obtain reliable results.

A natural choice of calibration technique for an HP8510B is the "Thru-Reflect-Line" (TRL) method which allows for measurements in nonstandard transmission media such as coplanar line. The transition between coaxial and coplanar line can, in principle, be accounted for in the calibration procedure. However, reproducible calibration standards in coplanar line are required. Reflections at connections between segments of coplanar line and uncertainties in the reflection standard lead to nonrepeatable results and unacceptably high errors. As a result, this method was not pursued further.

Our second technique was a simple 1-port measurement of a circuit which employs a direct connection from coaxial to coplanar line. This measurement of $|s_{11}|$ was made over a 1.5 GHz to 2.5 GHz frequency range and Fig. 2(a) shows the test arrangement. The abrupt transition from coaxial line to coplanar line was formed at the edge of the stycast substrate with a flange mount SMA connector. The measurement was taken with the reference plane of the backshort adjusted to coincide with the SMA connector.

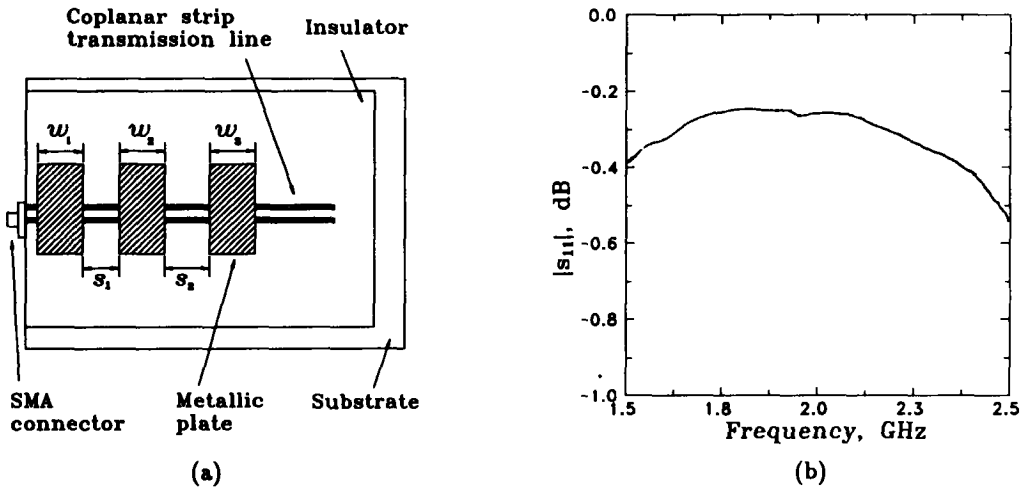


Fig. 2. Test circuit used for optimization (a) and plot of reflection coefficient (b) for optimized sliding backshort at the SMA connector on 204-Ω transmission line.

While this transition resulted in large unwanted reflections which increased the uncertainty of the measurement, the technique was useful because it allowed us to monitor the reflection coefficient for the backshort in real time as we optimized the dimensions of the backshort. The optimization was performed by systematically varying the length of the low and high impedance sections, the number of sections, the dimensions of the transmission line, and the thickness of the insulator in order to achieve the largest reflection of RF power. Good performance was obtained for a coplanar transmission line with 2.1 mm wide copper strips, separated by a 5.2 mm gap and mounted on a 6 mm thick stycast substrate with a dielectric constant of 4. The characteristic impedance of this transmission line and its effective dielectric constant were determined to be 204 Ω and 2.3 respectively [1]. A 0.025 mm thick sheet of mylar was used to insulate the transmission line from the sliding shorting plate. This noncontacting, 76 mm wide, 6 mm thick aluminum shorting plate had two rectangular holes in it with dimensions and spacing of $w_1 = 24.3$ mm, $s_1 = 19.4$ mm, $w_2 = 24.0$ mm, $s_2 = 23.0$ mm, and $w_3 = 24.4$ mm. This resulted in uncovered high impedance sections, s_1 and s_2, and covered low impedance sections, w_1, w_2 and w_3, which were each approximately $\lambda_g/4$ long on the coplanar line.

A plot of $|s_{11}|$ versus frequency is shown in Fig. 2(b). This optimized planar backshort produced $|s_{11}|$ better than −0.3 dB over a 20% bandwidth. That is a reflection of more than 90% of the power in the incident wave. The center frequency was determined to be 2 GHz, which agrees exactly with the design frequency.

We noted that the front edge of the sliding metallic plate was very close to the discontinuity at the SMA connector and it may have interacted with fringing fields around the flange. In addition, the reflection coefficient decreased when the backshort was moved $\lambda_g/2$ from the SMA connector in order to reduce interactions with fringing fields. The error caused by this unwanted interference, along with the lack of a transmission measurement, motivated the design of a third measurement technique.

Fig. 3 shows the system used for the third measurement technique. Here, the 204-Ω coplanar line was connected to the 50-Ω network analyzer inputs by means of two baluns of identical length. These baluns were made by gradually trimming the shield and teflon insulation from a semirigid, 3.5 mm wide coaxial line over approximately one wavelength at 2 GHz. This created a smooth transition to the coplanar line which minimized the power reflected at the connection. The return loss for these baluns is approximately −10 dB and the remaining undesired reflections from these transitions were gated out of the measurement using the low-pass time domain mode of the network analyzer. The frequency for this measurement was swept from 50 MHz to 20 GHz so that an accurate transformation between frequency and time could be made.

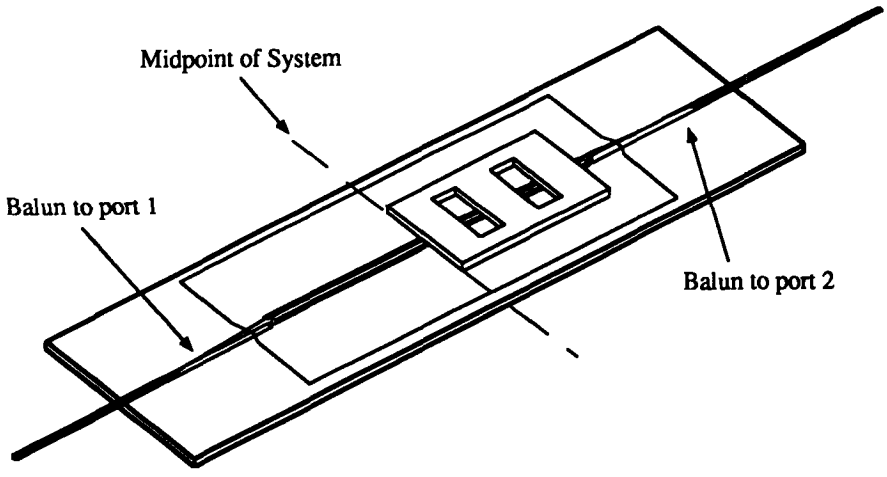


Fig. 3. Test system used to measure two-port scattering parameters for sliding backshort. The coaxial line baluns are tapered to create a gradual transition between coaxial and coplanar line which reduces measurement uncertainty due to unwanted reflections at the transition.

The full two-port scattering parameters for the system were measured under three different conditions. First, a reference measurement was made which would correspond to an ideal short. The baluns are identical in length and hence, the test model is symmetric about the midpoint of the coplanar line. Thus, the magnitude of the transmission measurement of this circuit with no short in place is equal to the reflection measurement with an ideal short at the midpoint. Reflection measurements for the sliding backshort were then made with the shorting plate arranged to reflect an incident wave from port 1 at the midpoint of the system and then, port 2. The values for $|s_{11}|$ from the first reflection measurement and $|s_{22}|$ from the second were normalized by dividing each by the values for $|s_{21}|$ and $|s_{12}|$ from the reference measurement, respectively. The two results were averaged to cancel the effect of any asymmetry in the system. Transmission measurements for the backshort were similarly normalized and averaged. The requirement of processing the measurement data, along with the large range of frequency, prohibited us form monitoring the 2 GHz reflection coefficient in real time for this measurement.

The results for the averaged, normalized $|s_{11}|$ and $|s_{21}|$ are shown in Fig. 4. The plot shows that $|s_{11}|$ was

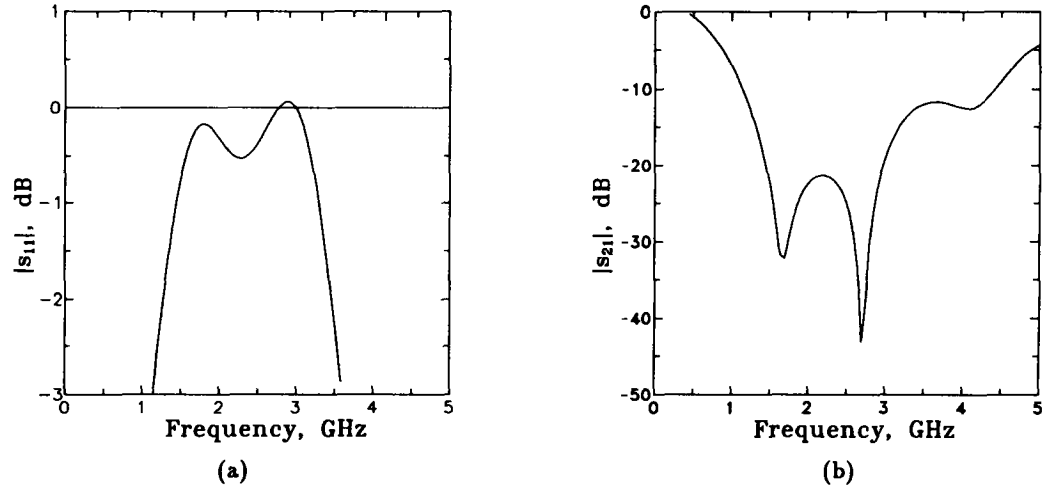


Fig. 4. Plot of measured reflection coefficient (a) and transmission coefficient (b) for optimized sliding backshort. This measurement was made using baluns for transitions.

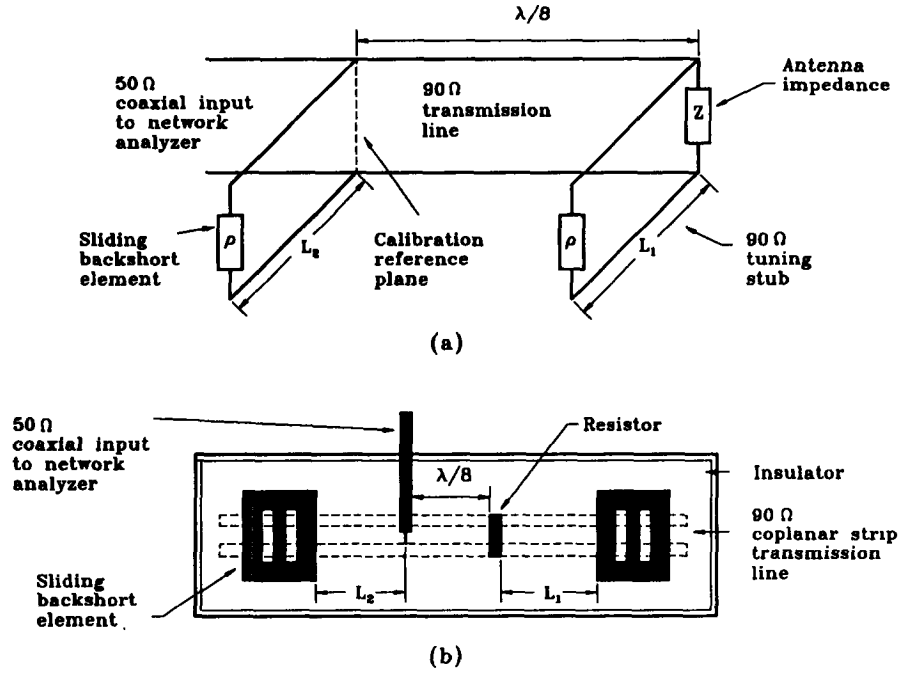


Fig. 5. Schematic diagram for equivalent double-shunt stub tuner circuit (a) and circuit arrangement used for measurements (b).

better than −0.5 dB over approximately an 80% bandwidth and the center frequency was slightly higher than 2 GHz. Over the peak located between 1.5 GHz and 2.5 GHz, $|s_{11}|$ is better than −0.3 dB over a 16% bandwidth and the center frequency is slightly lower than 2 GHz. This agrees well with the previous results shown in Fig. 2(b). Together, the plots of $|s_{11}|$ and $|s_{21}|$ appear to indicate that approximately 10% of the power for the incident wave is left unaccounted for, but the ± 0.2 dB uncertainty of our measurement is to large to verify this.

DOUBLE STUB TUNER

In order to demonstrate the tuning range accessible in a planar circuit, we built a double-shunt stub tuner which incorporates two sliding backshorts. This circuit serves as a low frequency model of a superconductor-insulator- superconductor (SIS) mixer used in millimeter wave radioastronomy applications [2,3]. Fig. 5(a) shows the equivalent circuit and Fig. 5(b) shows the circuit arrangement as measured. The characteristic impedance of the coplanar line is 90 Ω and the stub spacing is $\lambda_g/8$. A 90-Ω resistor was used to simulate a planar antenna impedance and a 3.5 mm wide semirigid coaxial probe was used to measure the range of impedance to which we could transform the resistor. The calibration reference plane for this measurement was set at end of the shield for the coaxial probe. Fig. 6 is a Smith chart, normalized to the characteristic impedance of the line, which shows the accessible impedance region at 2 GHz. The overlap for the tuning region and the impedance region necessary for matching SIS devices implies that a circuit of this type would be useful for this purpose. Variations in the shape of this tuning region can be achieved by changing the spacing between the tuning stubs.

The solid and dashed boundary lines in Fig. 6 show a comparison between the measured impedance range of the double stub tuner and a computer simulation of the circuit using *Puff* [4], respectively. The simulation agrees well with the measured data. The reflection coefficient of the shorting element used in the computer model was adjusted to fit the simulation to the measured Smith chart data. The resulting fitted $|s_{11}|$ for the backshort was −1.0 dB, which is similar to the −0.7 dB result that was measured for the

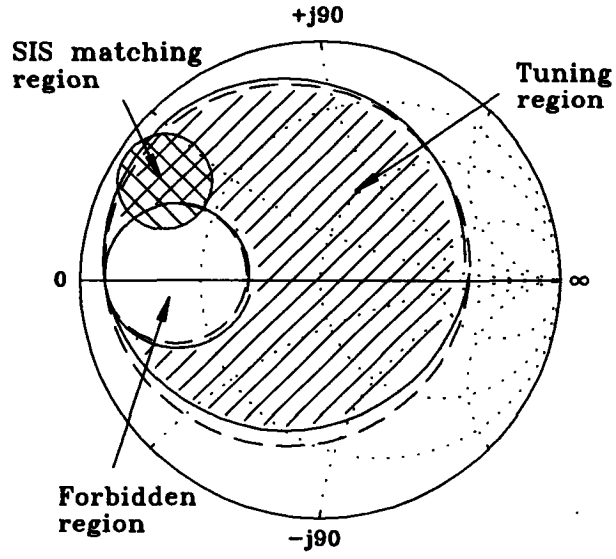


Fig. 6. Smith chart showing the measured (solid boundary) and fitted (dashed boundary) tuning region for double shunt stub tuner. The available tuning region covers the impedance region needed for matching to an SIS device.

same backshort, by itself, on a 90-Ω transmission line at 2 GHz. A phase shift of 4 degrees was also fitted at the coaxial probe transition in order to account for the calibration reference plane uncertainty.

CONCLUSION

We have demonstrated an approach for an adjustable planar backshort on coplanar strip transmission line. Test results from a low-frequency model indicate that the backshort can be used to create tuning stubs whose electrical length can be varied after fabrication. This noncontacting backshort with cascaded high and low impedance sections should also work on slotline, coplanar waveguide and possibly microstrip line. By using advanced micro-machining techniques [5,6], it should be possible to create adjustable impedance matching circuits at terahertz frequencies which would relax the design constraints for a wide range of planar integrated circuits.

ACKNOWLEDGEMENTS

We wish to thank Y-C. Tai for his help in keeping the design viable for terahertz scaling through micro-machining techniques. We also wish to thank O. Borić, M. A. Frerking, E. Kollberg, K. Potter, P. Siegel, and T. Tolmunen for valuable discussions. This work was supported in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration and the Innovative Science and Technology Office of the Strategic Defense Initiative Organization.

REFERENCE

[1] Y.T. Lo, S.W. Lee, *Antenna Handbook*, Van Nostrand Reinhold Co., N. Y., p. 28-35, 1988.

[2] J.R Tucker and M.J. Feldman, "Quantum Detection at Millimeter Wavelengths," *Reviews of Modern Physics*, vol. 57, no. 4, pp. 1055-1113, October 1985.

[3] Q. Hu, C A. Mears, P.L. Richards, and F.L. Lloyd, "MM Wave Quasioptical SIS Mixers," *IEEE Transactions on Magnetics*, vol. 25, no. 21, pp. 1380-1383, March 1989.

[4] S.W. Wedge, R. Compton, D. Rutledge, *Puff: Computer Aided Design for Microwave Integrated Circuits*, (published at Caltech, Pasadena, California), 1991.

[5] L-S. Fan, Y-C. Tai, and R.S. Muller, "Integrated Movable Micromechanical Structures for Sensors and Actuators", *IEEE Transactions on Electron Devices*, vol. 35, no. 6, pp. 724-730, June 1988.

[6] M. Mehregany, K.J. Gabriel, and W.S.N. Trimmer, "Integrated Fabrication of Polysilicon Mechanisms," *IEEE Transactions on Electron Devices*, vol. 35, no. 6, pp. 719-723, June 1988.

MATERIALS SCIENCE

(Session D5/Room A2)

Thursday December 5, 1991

- **Passive Chlorophyll Detector**
- **Commercial Application of Thermal Protection System Technology**
- **Oxynitride Glass Fibers**
- **Commercial Applications of Advanced Photovoltaic Technologies**

PRECEDING PAGE BLANK NOT FILMED

PASSIVE CHLOROPHYLL DETECTOR

Leonard A. Haslim
Research Scientist
NASA Ames Research Center
M.S. 237-11
Moffett Field, CA 94035-1000
(415) 604-6575

DOCUMENTATION OF THIS PAPER WAS NOT PROVIDED IN TIME FOR INCLUSION IN THESE PROCEEDINGS. FOR FURTHER INFORMATION, PLEASE DIRECT ALL INQUIRIES TO THE PERSON LISTED ABOVE.

PRECEDING PAGE BLANK NOT FILMED

N92-22705

COMMERCIAL APPLICATION OF THERMAL PROTECTION SYSTEM TECHNOLOGY

Gordon L. Dyer
Technology Transfer Officer
George C. Marshall Space Flight Center
Martin Marietta Manned Space Systems
P. O. Box 29304
New Orleans, LA 70189

INTRODUCTION

This paper focuses on the thermal protection system process technology used in the manufacture of the External Tank for the Space Shuttle system and how that technology is applied by private business to create new products, new markets and new American jobs.

The term "technology transfer" means different things to different people and has become one of the buzz words of the 1980s and 1990s. In the context of this paper, "technology transfer" is defined as a means of transferring technologies developed by NASA's prime contractors to public and private sector industries.

Background - Manned Space Systems and Technology Transfer

Despite the tens of thousands of spinoff products and processes from NASA, the majority of small to mid-sized businesses in the United States do not understand the term or concept of technology transfer. In fact, you might say that technology transfer is one of the best kept secrets in America.

Prime contractors like Martin Marietta Manned Space Systems have always had the contractual requirement to report all new technology developments to the contracting agency. In October, 1989, Martin Marietta Manned Space Systems' contract was expanded to expedite the movement of technologies from the "technology closet" to the market place.

Through this new contract clause, Martin Marietta Manned Space Systems has set out to familiarize the public and private sectors with how easy it is to take advantage of Federally-developed technologies for commercial usages through the "NASA/Martin Marietta Manned Space Systems Technology Transfer Program". As a result of these efforts, four commercial applications of External Tank-developed technology have transpired since October 1989. Two of these applications will be discussed in this paper.

Background - Federal Legislation

Although "technology transfer" is one of America's best kept secrets, it shouldn't be, according to the legislation passed to support the effort. Technology transfer was one of the major considerations of the National Aeronautics and Space Administration when it was created. In Congress' first effort to show a commitment to technology transfer, the National Aeronautics and Space Act of 1958, as amended, provided a clear mandate for the NASA to have an active Technology Transfer Program. However, simply mandating a Technology Transfer Program fell short of getting American business interested in developing

NASA-derived technology. The lack of interest by business was due to the fact that newly developed technology was considered public domain.

Later attempts by Congress in 1980 to foster the commercialization of Federally-developed technology resulted in two laws being passed. For the first time, the Bayh-Dole Act allowed the government to grant exclusive commercial sector rights to Federally-developed technology to the contractor who developed the technology with federal funding. By allowing this technology to be exempt from the public domain, this law removed a major stumbling block for private industry investment.

The second law, the Stevenson-Wydler Act authorized Federal laboratory personnel to share government-funded technology with the private sector. Until this law was passed, other than NASA, Technology Transfer was not part of the Federal laboratories mission.

Still not seeing the kind of results it was looking for, the Congress enacted the Federal Technology Transfer Act of 1986 in response to the increasingly tough international economic competition facing the United States from Japan, the Pacific Rim nations and the industrialized nations of Europe. The Federal Technology Transfer Act along with President Reagan's multi-phased program for improving access to Federally-funded research contained in Executive Order 12591, charged Federal agencies with linking their Federally-developed technologies with the private sector.

NASA, with its long-standing history in the Technology Transfer arena and in the spirit of the Federal Technology Transfer Act, decided through the Marshall Space Flight Center (MSFC) to increase their technology transfer efforts by allowing its prime contractors to implement complementary programs. This pilot program being conducted by Martin Marietta Manned Space Systems is helping to expand the network for transferring NASA/contractor-developed technologies.

Background - External Tank

Martin Marietta Manned Space Systems, Michoud Assembly Facility, New Orleans, Louisiana manufactures the External Tank for the Space Shuttle system. The Space Shuttle system is comprised of four major elements: The airplane-like orbiter, its external propellant fuel tank, and twin solid rocket boosters. The Space Shuttle system's largest element, the External Tank, is the structural backbone of the vehicle. It provides liquid cryogenic propellants to the orbiter's three main engines and absorbs the 6.7 million pounds of thrust exerted by the solid rocket boosters and main engines during launch.

To manufacture an External Tank that is 154 feet long, nearly 28 feet in diameter and has a propellant capacity of 528,600 gallons (1.6 million pounds), numerous technologies must come into play.

To keep the liquid hydrogen and liquid oxygen in the External Tank prior to launch, at -423°F and -297°F respectively an excellent thermal protection system is required. A polyisocyanurate foam was chosen to perform the task. The External Tank is covered is covered over most of its surface with a layer of the foam about 1 inch to 1 1/2 inches thick. In addition to insulating the cryogenic liquid contents of the tank from its exterior environment and keeping these propellants from boiling off or causing ice to form on the exterior of the tank, the foam insulation also protects the tank's aluminum skin from the extreme heat caused by aerodynamic heating during the ascent portion of the mission. The foams and application processes developed to allow the External Tank to survive the extreme conditions it encounters are now finding their way into the commercial sector, helping companies create or improve their products.

MARTIN MARIETTA MANNED SPACE SYSTEMS TECHNOLOGY TRANSFER PROGRAM

Manned Space Systems' first experience with transferring space-developed technology was a project referred to as the "Childrens Lunchbox Meals". Eleven million children do not have access to proper lunch time meals at school. With that fact in mind, Dr. Blackwell founded SouthPointe Products in Birmingham, Alabama, to manufacture and market a unique food product. SouthPointe Products offers parents a healthy alternative for their children's diets with a product that is nutritious, convenient to prepare, and that can be heated at home and retain its warmth for several hours until the child is ready to have lunch. The meals will be packed in thermoformed, high barrier plastic single serve trays with non-foil, easy peelable lidding. Some meals will be packaged for the child to heat and eat immediately; others will be provided with an insulated "meal mitt" or "cocoon" so that they can be heated and stored in the child's lunch box for several hours prior to eating. Dr. Blackwell's problem was to develop a "meal mitt" or "cocoon" that would insulate the meal and maintain the temperature of the meal at approximately 110°F for four to five hours. She contacted the Marshall Space Flight Center with her technical problem in the field of thermal materials. This problem was forwarded to Martin Marietta Manned Space Systems since one of the major areas of our expertise is the field of thermal protection technology developed for the External Tank.

After Manned Space Systems engineers established that this problem could be solved using thermal protection materials and processes developed under the External Tank contract, an applications engineering project was initiated with SouthPointe Products. This project included (1) analysis and testing to prove the concept of using foam as the "cocoon" material, (2) vendor surveys to determine production feasibility/costs to manufacture Manned Space Systems' concept, and (3) development of prototype "cocoons" for SouthPointe Products.

The results of this project demonstrated that it was feasible to use urethane foams as insulation for this microwaveable meal. A 0.5 inch thick foam shell would be required to keep the meal warm (~110°F) for four to five hours (Ref. Figures 1 & 2). Dr. Blackwell's production cost will vary with the vendor making the shells and the foam being used.

At the completion of this applications engineering project all patent and intellectual rights were assigned to SouthPointe Products. NASA, however, by law always retains a non-exclusive, non-transferrable, irrevocable, paid-up license for any technology developed with government funding.

Currently, SouthPointe Products has a patent pending on the foam "cocoon" concept and is working with an Atlanta firm to produce and market Dr. Blackwell's quality lunchbox meals for kids, coined "Power Packs".

Another example of an applications engineering project that will lead to commercial application of thermal protection system technology is the thermal curtain project. United Service Equipment Company (USECO) of Murfreesboro, Tennessee manufacture and markets the durable UNITRAY® cart, the original and completely proven single-tray heated and refrigerated food cart. Each of the double tray compartments is divided into two sections, one heated and the other refrigerated. It is this unique system that permits the tray to be assembled as a completed meal in the central kitchen, under the supervision of the dietitian, and delivered directly to the patient. Both hot and cold foods, side by side on the same tray, arrive at bedside appetizingly fresh because the UNITRAY® is designed not only to transport the trays but also to maintain food temperatures.

USECO has manufactured the UNITRAY® cart since the mid-1960s. From that time through the mid-1970s it was a very big seller for them. In the late 1970s to the early 1980s, the UNITRAY® started to lose its market appeal. Other manufacturers had introduced several products similar to the UNITRAY®, but with superior insulating properties. This new marketplace competition led USECO to contact NASA through the State of Tennessee Department of Economic and Community Development in search of technologies that could replace the neoprene divider system they were using at that time.

Because of Martin Marietta Manned Space Systems' expertise in thermal protection systems, the George C. Marshall Space Flight Center asked that we review this problem.

A team of engineers representing materials engineering, thermal engineering, mechanical design engineering and advanced manufacturing technology engineering were put together to review the USECO UNITRAY® food cart current thermal barrier. Several concepts were developed and presented to USECO for their opinion and consideration. Upon USECO agreeing with a proposed concept to utilize thermal protection materials to solve their problem, an applications engineering project was begun with USECO.

The objective of this program was to develop a foam-based thermal curtain system concept (Ref. Figures 3 & 4) which provides an insulating barrier between hot and cold chambers capable to maintaining fixed temperature conditions at 80-100 percent relative humidity for a given period of time. Test conditions would include maintaining contrasting temperatures of 250-275°F and 34-40°F for 2 hours.

The approach to conducting this project include: (1) The development of engineering drawings, (2) development of prototype molds and prototype parts, (3) in-house analysis and testing to prove the concept, and (4) mocking up the UNITRAY® cart with prototype parts and performing an environmental acceptance test as the final proof of concept.

USECO plans on reintroducing the UNITRAY® cart with the new foam barrier systems in 1992. As with the lunchbox project, Martin Marietta Manned Space Systems disclosed all reportable items to NASA pursuant to NASA FAR supplement clause 18-52.227-70, "New Technology" and at the same time requesting that NASA waive its rights to the "UNITRAY® Delivery Cart Thermal Curtain." Upon the issuance of this waiver, Martin Marietta Manned Space Systems will assign its rights to the United Service Equipment Company who will file patent applications on the thermal barrier.

CONCLUSION

With results like these, it is easy to understand why "Technology Transfer" can and does work for the American people. The projects cited in this paper deal with thermal protection systems technology. However, these are just two of the several applications engineering projects Martin Marietta Manned Space Systems has been involved in since October 1989. Manned Space Systems also has expertise in the fields of advanced manufacturing, advanced inspection systems, and advanced materials.

The NASA/Martin Marietta Manned Space Systems Technology Transfer Program is a positive way to take advantage of the technology being developed with your tax dollars, and a way of putting the United States back in the driver's seat as the world leader in manufacturing.

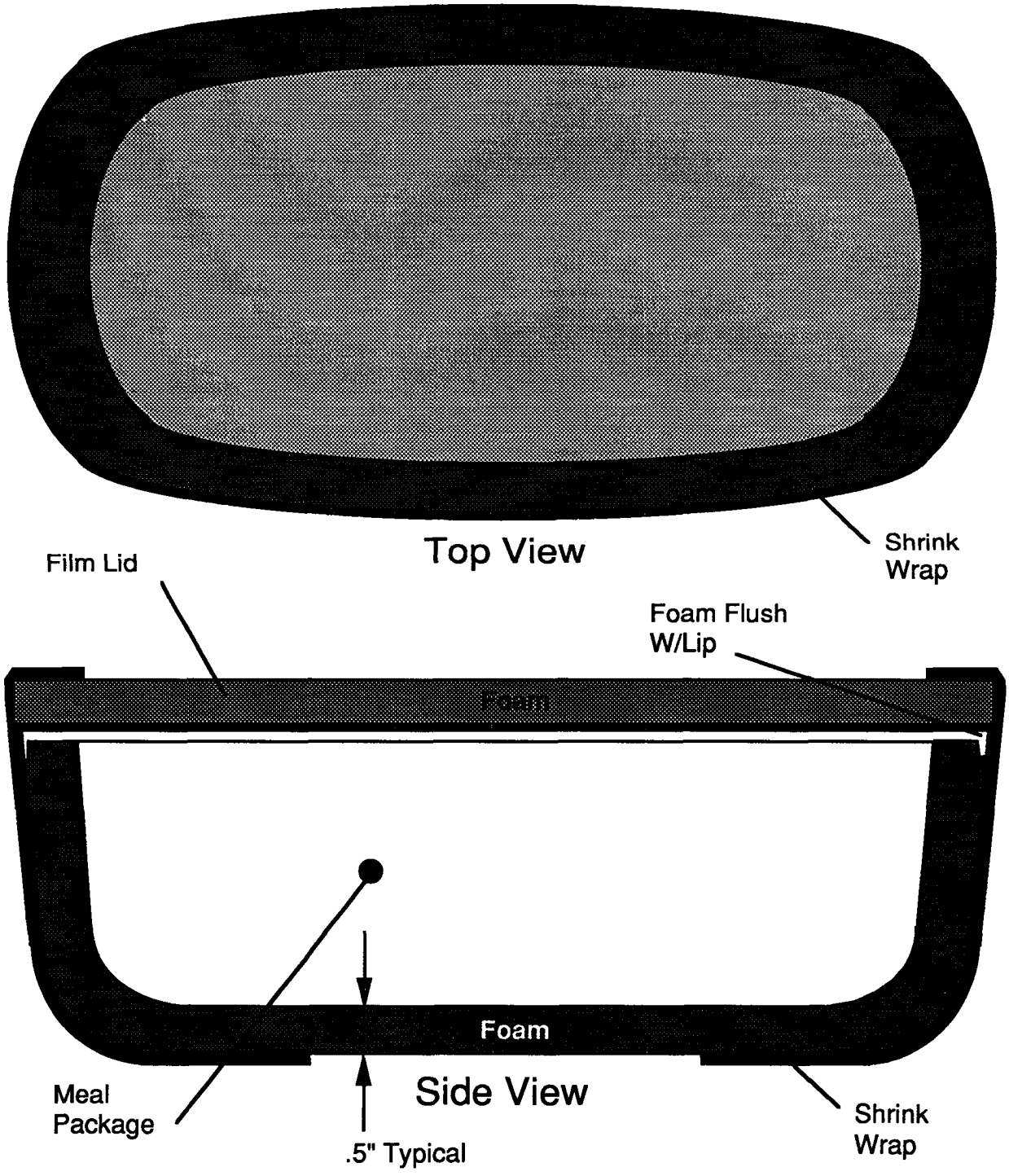


Fig. 1 SouthPointe Products' Children Lunchbox

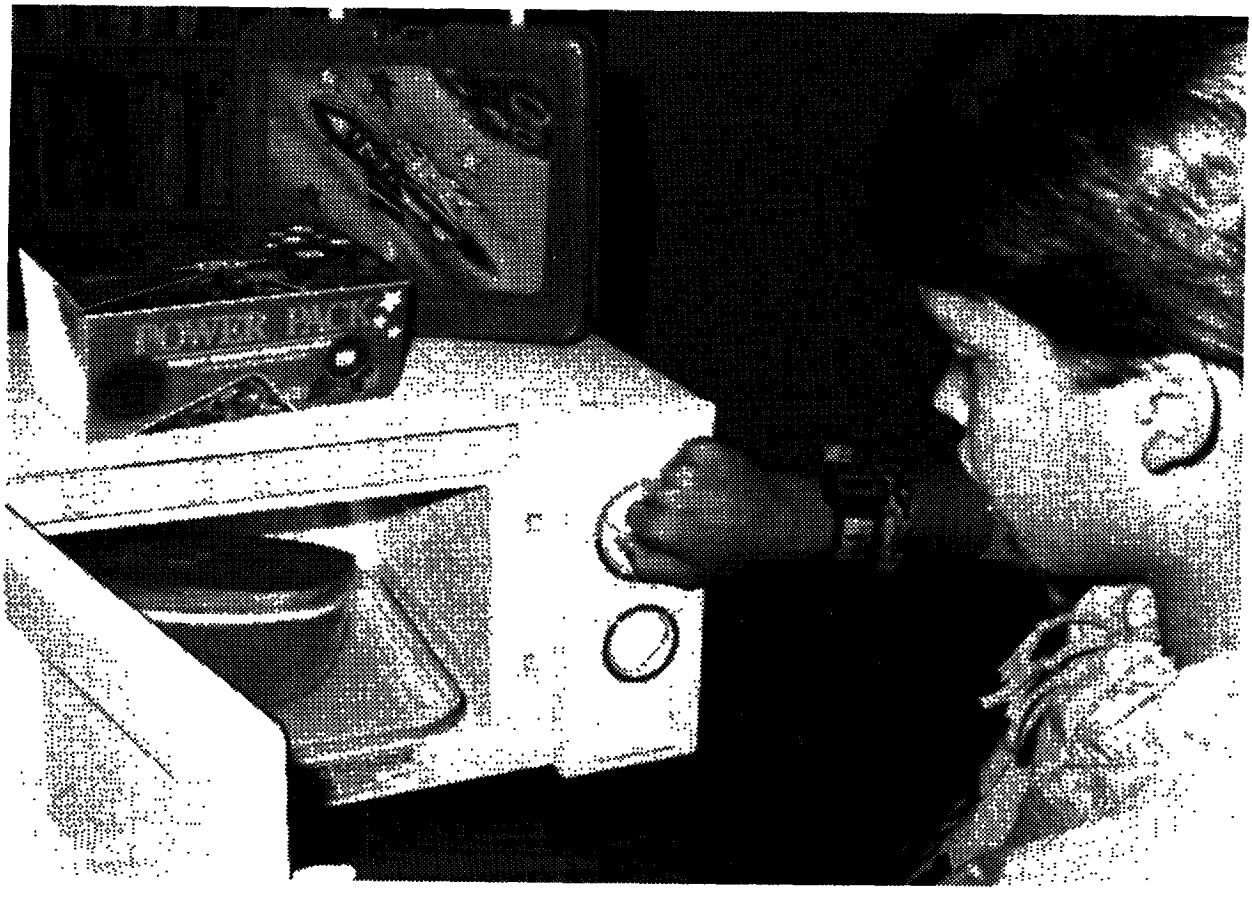
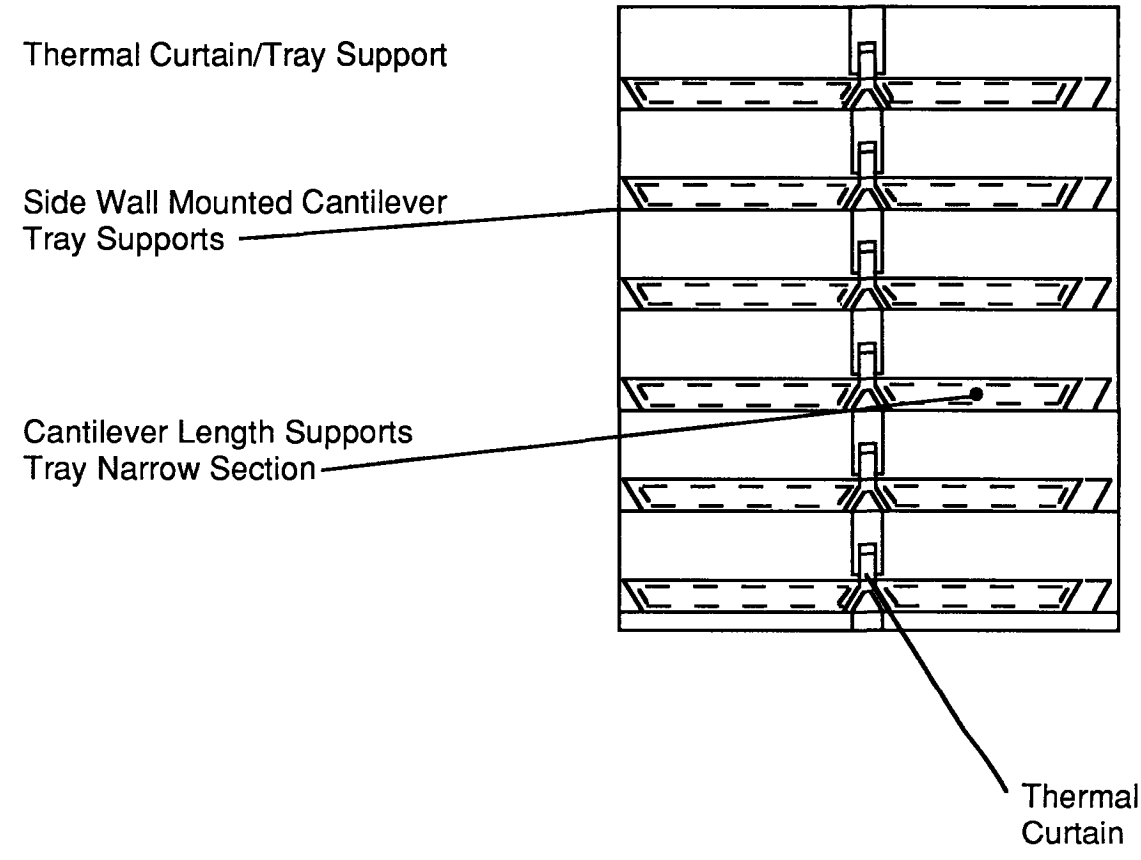


Fig. 2 Children's Lunchbox Being Heated in Microwave

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



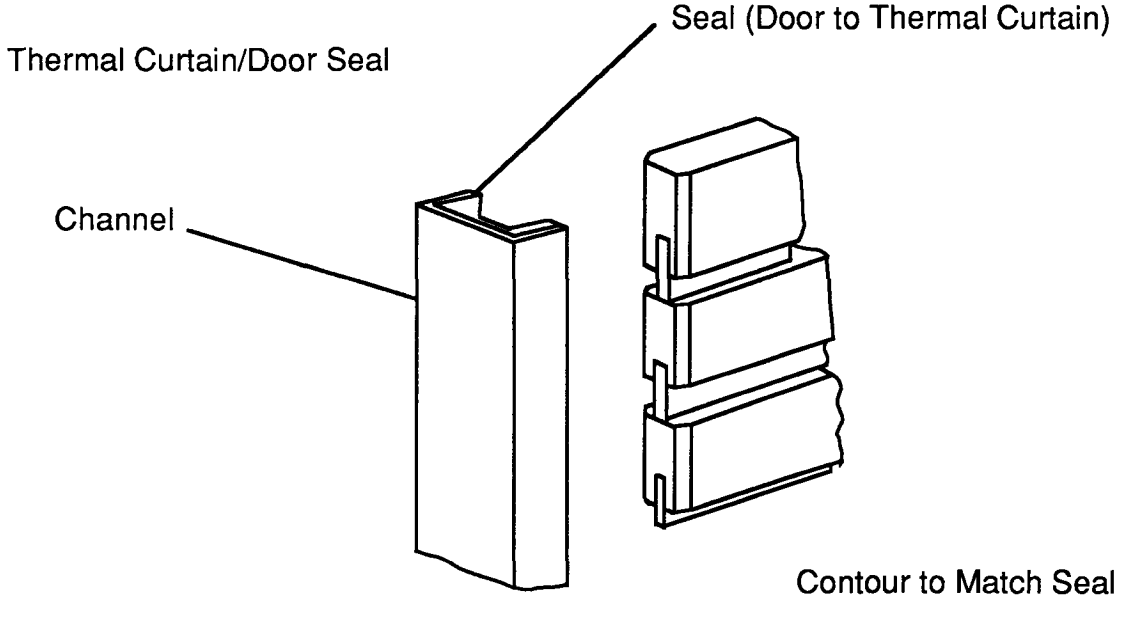


Fig. 3 Thermal Curtain

Fig. 4 Thermal Curtain New Foam Concept and Current Divider System

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

N92-22706

OXYNITRIDE GLASS FIBERS

Parimal J. Patel, Donald R. Messier and R.E. Rich
Research Ceramic Engineer
ATTN: SLCMT-EMC
Army Laboratory Command
Materials Technology Laboratory
Watertown, Massachusetts 02171-0001

ABSTRACT

Research at the Army Materials Technology Laboratory (AMTL) and elsewhere has shown that many glass properties including elastic modulus, hardness, and corrosion resistance are improved markedly by the substitution of nitrogen for oxygen in the glass structure. Oxynitride glasses therefore offer exciting opportunities for making high modulus, high strength glass fibers. Discussed in this paper are processes for making oxynitride glasses and fibers of glass compositions similar to commercial oxide glasses, but with considerably enhanced properties. We have made glasses with elastic moduli as high as 140 GPa (20 Msi) and fibers with moduli of 120 GPa (17 Msi) and tensile strengths up to 2900 MPa (420 ksi). AMTL holds a U.S. patent on oxynitride glass fibers, and this presentation discusses a unique process for drawing for drawing small diameter (10 *um*) oxynitride glass fibers at high drawing rates (1500 m/min). Fibers are drawn through a nozzle from molten glass in a molybdenum crucible at 1550°C. The crucible is situated in a furnace chamber in flowing nitrogen, and the fiber is wound in air outside of the chamber, making the process straightforward and commercially feasible. Strengths have been considerably improved by improving glass quality to minimize internal defects. Though the fiber strengths have been comparable with oxide fibers, work is currently in progress to further improve the elastic modulus and strength of the fibers. The high elastic modulus of oxynitride glasses indicate their potential for making fibers with tensile strengths surpassing any oxide glass fibers, and we hope to realize that potential in the near future.

INTRODUCTION

The incorporation of nitrogen into an oxide glass network has been shown to improve several properties of the glass. Increasing nitrogen content leads to increases in glass transition temperature, viscosity, density, hardness, corrosion resistance, and elastic modulus.[1] For applications of high-performance resin matrix composites, the elastic modulus and strength need to be as high as possible and the density as low as possible. Presently, the continuous fibers used for advanced composites have been s-glass and e-glass fibers. S-glass fibers are from the Mg-Si-Al-O system while e-glass fibers have a Ca-Si-Al-O system. Work done at the U.S. Army Materials Technology Laboratory has focused on the processing of bulk oxynitride glass that are analogous to the s-glass composition for the purpose of drawing the glass into small diameter, high strength, high elastic modulus fibers for use in high-performance resin matrix composites. Bulk glasses have been produced with elastic modulus values of up 140 GPa. Earlier research at MTL as shown that the fibers retain all the properties of the bulk glass.

The oxynitride glass fiber program first focused on oxynitride glasses that had improved properties over oxide glasses. Once the bulk glasses were produced, they were attempted to be drawn into fibers. The criteria for the glass was to have high modulus, low density, and be fiberized easily. The third factor is the crucial one. Bulk oxynitride glasses with superior properties have been produced at M.T.L. However, drawing thin glass fibers from these compositions has not always been easy. It is this qualifying factor that determines which compositions can be studied. After further investigation,two glass compositions were chosen for examination. Both had compositions similar to commercial s-glass. However, one had 2.72 atomic percent nitrogen and the other had 4.13 atomic percent nitrogen. These fibers had moduli that were greater than the oxide glass fibers. High modulus is the most important criteria for composites and it is in this aspect that the oxynitride glass fibers are superior. However, the tensile strengths of the oxynitride glass fibers were too low, especially when considering their potential. The effort was then put through to

improve the tensile strength of the fibers. The following paper describes that process and the accomplishments made in the fiber properties.

PROCEDURE

Glass:

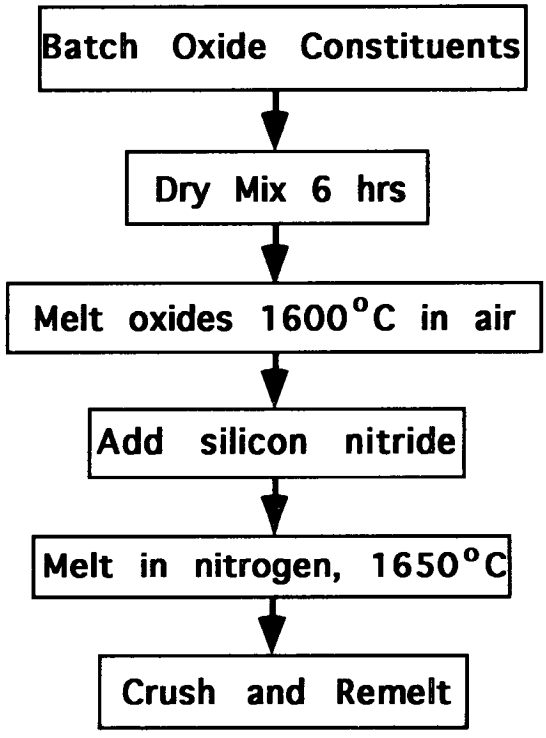
Batches were prepared using the compositions listed in Table I. To optimize the processing of the glasses, two base compositions were used so that comparison would be possible. As shown, one composition, NS8, was a MgO-Al_2O_3-SiO_2-Si_3N_4 composition. This glass contains 2.72 mole % nitrogen incorporated into the glass network. The other glass composition, NS 7.5, had 4.13 mole % nitrogen of the same batch materials. These two compositions were used for their good drawing characteristics thus allowing the optimization of the processing of high strength fibers.

Table I: Oxynitride Glass Base Composition

Species	NS8 (2.71 at. % N) Wt. Percent	NS7.5 (4.13 at. % N) Wt. Percent
MgO	10.13	10.20
Al_2O_3	25.34	25.51
SiO_2	59.79	57.13
Si_3N_4	4.74	7.16

Figure I outlines the process that has been developed for the oxynitride glasses. The oxide constituent of the glasses (MgO-Al_2O_3-SiO_2) were dry mixed for six hours. Wet mixing with acetone was found undesirable due to iron contamination of the acetone. The oxides were melted in air at 1600°C for two hours and quenched in water. The oxide glass was then mixed with silicon nitride (Union Carbide # 1929) and placed in a molybdenum foil crucible. The crucible was placed into a alumina crucible and loaded into a water-cooled, tungsten-mesh resistance furnace. The sample was melted at 1650°C in nitrogen at a dwell time of 2 hours and a subsequent furnace quench of 50°C per minute. The sample was removed, crushed, and remelted at the identical thermal cycle used for the first melt. At this point, the glass was ready to be characterized and drawn into fibers.

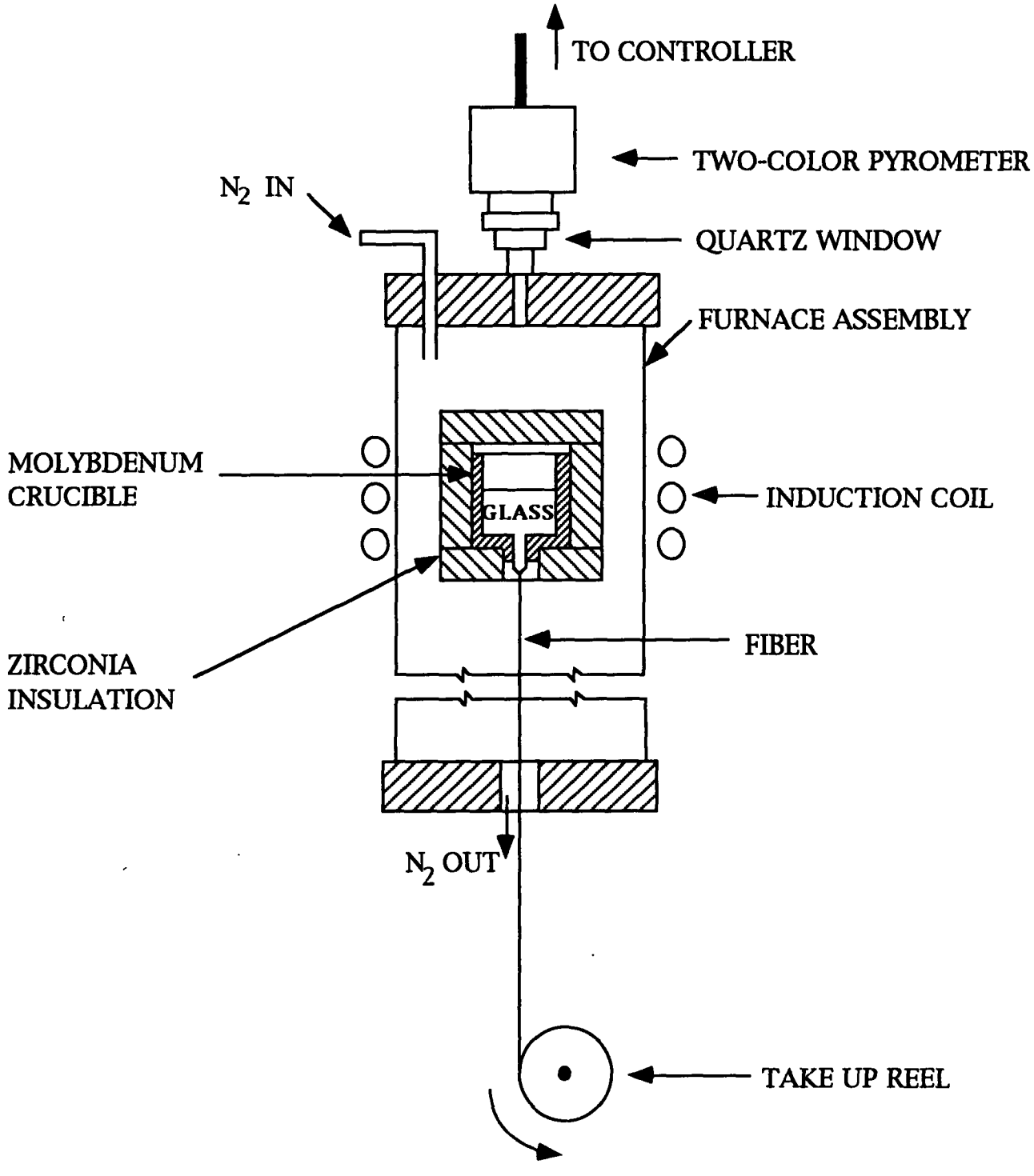
Figure I: Oxynitride Glass Fabrication Process



Fibers

Once the glass was produced, it was ready to be drawn into fibers. The fiber drawing process used is very similar to commercial processing of continuous glass fibers. Figure II is a schematic representation of the fiber drawing process. The glass is placed in a molybdenum single nozzle crucible. The furnace assembly is located in a furnace chamber with flowing nitrogen. After ample time for the nitrogen flow to fill the tube the furnace is inductively heated to 1500°C- 1600°C, depending on composition. Once the temperature has stabilized, a nut is removed from the bottom fixture of the silica glass tube. A thin alumina rod is then fed through to the bushing where a glob of glass has accumulated at the nozzle. The glass adheres to the cooler alumina rod allowing the glass to thin out into a fiber as the alumina rod is pulled through the bottom of the hole. The fiber is broken off the rod and pulled manually to the winder where it is taped on. As one can see from Figure II, only the glass is exposed to the nitrogen atmosphere. The fiber is pulled in air. This design allows the process to be straightforward and commercially feasible. The winder motor speed is increased until it reaches a speed of 2400 r.p.m. or a draw rate of 1500 meters per minute. The fiber diameter, at this speed, is roughly between 8 and 12 microns. The only limitation for the fiber diameter is the drum winder speed. Presently, 2400 r.p.m. is the maximum speed of the winder. At this point, the fiber can be drawn continuously until the glass supply has diminished. However, for the purpose of characterization of the fibers, the draw is interrupted several times for sampling.

Figure II: Schematic Representation of Oxynitride Glass Fiber-Drawing System



Characterization

Prior to the fiber draw, the bulk glass was characterized using optical techniques. The glass was observed through a microscope with transmitted light to determine the presence of any crystalline material or unmelted batch. If the glass was determined to be of good quality, the glass would drawn into fibers. The density of the glass was also calculated using Archimede's method. Microhardness was also measured on polished sections using a Knoop indenter and a .98 N load.

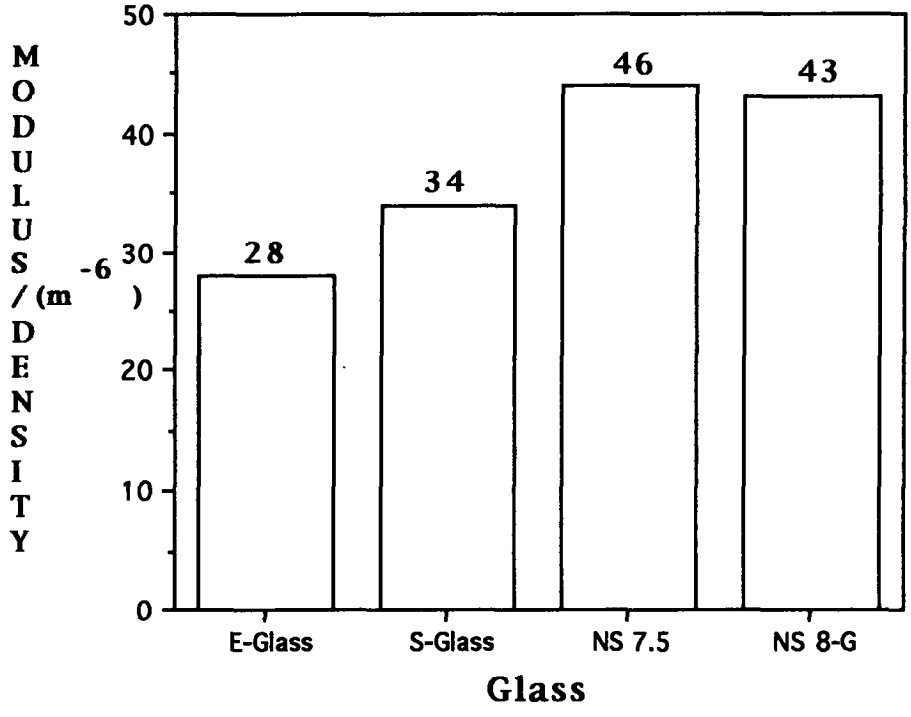
During the draw of the fibers, the fiber is broken several times for testing. Sampling is a crucial step in the characterization of the glass fibers. The fibers that are drawn do not have any sizing or any protective coating on them due to the experimental nature of the work. They are thus easily damaged and great care must be taken in their handling. The small diameters of the fibers further increased the difficulty in handling. The fibers were pulled off the line of the draw once the proper draw speed had been attained. The fibers were taped onto an I-shaped holder. A tape of seven fiber mounts with a gage length of 30 mm were placed behind the fiber. The fiber was taped to the mounts leaving the fiber within the slot untouched. Once all the fibers have been mounted, they were cemented at ends of the cut out slots. After the cement had dried, the mounts were cut off the tape and broken in tension. The fibers were broken at a crosshead speed of 1 mm per minute. The diameter of the fibers and the force required to break the fibers were recorded for future calculation of the tensile strength. Once the fibers were broken, if the fiber did not shatter, the fracture surface was studied using a S.E.M. Surface fracture analysis was performed on all fiber draws to determine the fracture origin of the fiber. In addition to the fiber tensile strength, the elastic modulus of the fibers was also determined via a sonic modulus instrument that measured the speed of sound through the fibers.

RESULTS

As cited before, oxynitride glasses show considerable potential for advance composites. There is a marked improvement in the properties over oxide glass fiber counterparts. However, the tensile strength was relatively weak. One focus of work at the Materials Technology Laboratory was to produce the highest modulus, highest strength, lowest density oxynitride glass fibers possible. Using the procedure described earlier, great strides have been made in attaining that goal.

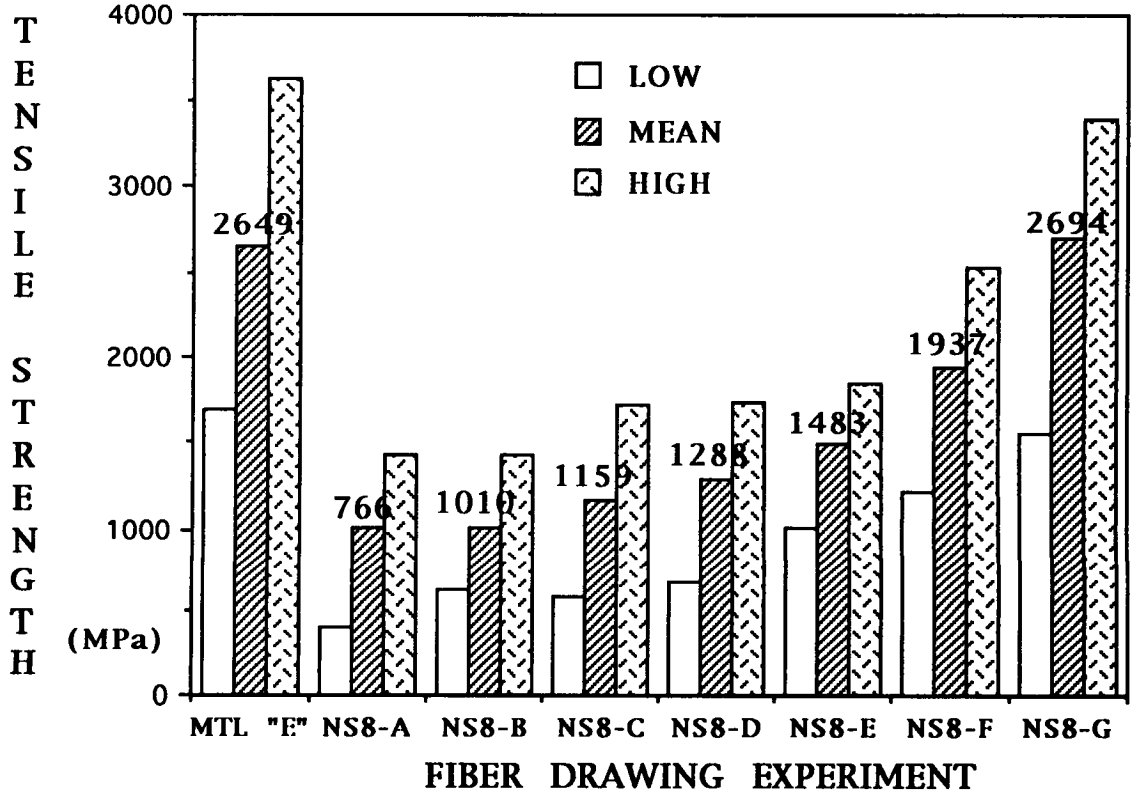
One of the important properties of the oxynitride glasses is the high elastic modulus, an important property necessary for advanced structural resin matrix composites. For aerospace application, the fibers also need to be of the lowest possible density. The modulus of the commercial e-glass with a density of 2570 kg/m^3 is 73 GPa.[2] The density and modulus of s-glass is 2480 kg/m^3 and 85 GPa3, respectively. An NS8 glass has been found to have a modulus of 105 GPa with a density of 2500 kg/m^3. The NS7.5 glass ,with more nitrogen, has a modulus of 112 GPa. Further examination into additives has led to glasses with modulus values of 120 GPa at a density of 2700 kg/m^3. Due to the the importance of density, one convention of looking at the modulus is the Elastic Modulus to Density Ratio (E/ρ). As seen on Figure III, the glasses made at M.T.L. have shown an improvement over conventional oxide glasses. The e-glass and s-glass had ratios of 28 m^{-6} and 34 m^{-6}, respectively. The oxynitride glass fibers displayed higher ratios. The NS 7.5 has a modulus to density ratio of 46 m^{-6}. NS8-G has an elastic modulus of modulus/density ratio of 43 m^{-6} compared to 28 m^{-6} of e glass and 34 m^{-6} of s glass. This glass also had tensile strengths comparable to oxide glass fibers.

Figure III: Modulus/Density Comparison



The tensile strength for the fiber is dependent on the composition as well as the processing. The base compositions, NS 8 and NS 7.5, were utilized for their good drawing properties and elastic moduli. Processing of the glass was maximized using these glass compositions. Figure IV illustrates the fiber tensile strength increases incurred due to the process changes implemented. As the process improvements have been made , the tensile strengths have become greater and greater. The MTL "E" glass refers to commercial e-glass marbles that were acquired from PPG Industries. These marbles were drawn in our fiber drawing apparatus as a standard of the fiber drawing and fiber testing techniques. The fibers tested a mean of 2649 MPa which were 15 to 20 percent weaker than industry standard. This is due to handling of the glass fibers. The tensile strength goal for the oxynitride glass fibers was to reach the strengths of commercial e-glass. The best attempt has been the NS 8-G glass fiber whose strength surpassed the MTL "E" with a mean strength of 2694 MPa. The highest tested value for the NS 8-G batch was 3320 MPa. This composition, as previously mentioned, had an elastic modulus of 121 GPa with a density of 2700 kg/m^3. This indicates the potential of higher strengths from tighter processing control. It is the first oxynitride glass fibers drawn at AMTL that couples the previous high modulus values with high tensile strength indicating the real potential for use in high performance resin matrix composites.

Figure IV:Mg-Si-Al-O-N Glass Fiber Strength Increases via Process Improvements



Future Work:

The oxynitride glass fibers produced thus far have already surpassed the oxide fibers with respect to elastic modulus and specific modulus. They have shown the potential for higher tensile strengths than their oxide counterparts. However, the high elastic modulus of the oxynitride glasses can still be further increased with incorporation of higher levels of nitrogen. Concurrently, the potential for higher strengths can also be improved via further process improvements. The goal is to determine the highest elastic modulus glass fibers producible with low density and high tensile strength and subsequently produce composite materials using the fibers.

REFERENCES

[1] D.R. Messier, R.P. Gleisner, R.E. Rich, "Yttrium-Silicon-Aluminum Oxynitride Glass Fibers," J. Am. Ceram. Soc., 72 [11] pp. 2183-86, (1986).

[2] P.F. Auborg and W.W. Wolf, Advances in Ceramics, 18, pp. 51-63, (1986)

[3] R.F. Lowrie, Modern Composite Materials, pp. 270-323, Addison-Wesley (1967)

N92-22707

COMMERCIAL APPLICATIONS OF NEW PHOTOVOLTAIC TECHNOLOGIES

R. McConnell
Technology Transfer Manager
National Renewable Energy Laboratory
Golden, CO 80401

ABSTRACT

The National Renewable Energy Laboratory (NREL) has directed and managed photovoltaic (PV) research and development (R&D) activities for the Department of Energy for more than 13 years. The NREL budget for these activities is almost $33 million for Fiscal Year 1991. With the world's increasing concern for the environment and the United States' renewed apprehension over secure and adequate energy supplies, the use of semiconducting materials for the direct conversion of sunlight to electricity--photovoltaics--is an excellent example of government-supported high technology ready for further development by U.S. companies. This paper will describe some new PV technologies and their research progress, some commercial applications of PV, and NREL's technology transfer activities for helping U.S. industry in its efforts to bring new products or services to the marketplace.

INTRODUCTION

The Solar Energy Research Institute (now the National Renewable Energy Laboratory) was established by Congress in 1974 as the nation's primary center for solar energy R&D. With almost 500 staff members, NREL is involved in scientific and engineering activities ranging from basic materials science, applied research directed towards performance improvements or cost reductions, and systems engineering directed toward project design and evaluation. Approximately 80 NREL staff are directly involved in PV research. Of the $33 million allocated for NREL's PV research during Fiscal Year 1991, over half was for subcontracted research by universities and private companies. Much of the subcontracted research with industry is cost-shared; that is, a company provides funds of its own to add to NREL subcontract funds. NREL's cost-shared subcontracted research with private companies overcomes many technology transfer barriers because industry researchers are directly supported for further development of PV technology.

NEW PV TECHNOLOGIES

Some of the PV technologies new to the marketplace are not at all new to PV researchers. NREL PV researchers have worked on these technologies since the late 1970s. The crystalline silicon PV technology was discovered in the mid-1950s and is still the most widely sold PV technology, principally because of its high efficiency and long-term stability. The new technologies, however, have the potential of being cheaper than crystalline silicon. Their potential for lower cost arises from significantly lower material requirements, lower energy processing, or higher volume production capabilities. Additional R&D, for example, to improve conversion efficiencies, is required for the new technologies to reach the marketplace. The new PV technologies in NREL's PV program are amorphous silicon thin films, polycrystalline thin films, III-V PV devices made from elements in columns III and V of the periodic table, and new approaches for using crystalline silicon.

AMORPHOUS SILICON THIN FILMS

Amorphous silicon is a disordered material without the crystalline structure of the silicon used in the semiconductor industry. Figure 1 schematically shows the atoms in amorphous silicon material (1). However, amorphous silicon absorbs solar radiation much more efficiently than does crystalline silicon, with dramatic implications for the amount of material needed and the future cost of production of the fully developed technology. Approximately 200 times less silicon is required for amorphous silicon PV devices than is needed for crystalline silicon devices. Amorphous silicon conversion efficiencies are respectable--more than 12% total area efficiency for small laboratory devices. An important research issue for amorphous silicon has been a 10%-30% decline in the conversion efficiency of production devices when they are exposed to sunlight.

POLYCRYSTALLINE THIN FILMS

Polycrystalline thin films of copper indium diselenide ($CuInSe_2$, abbreviated CIS) or cadmium telluride (CdTe) also absorb solar radiation much more efficiently than single-crystal silicon does. Again, small amounts (thin films) of material are needed so that future production costs for polycrystalline thin-film devices are expected to be lower than those for crystalline silicon PV devices. Like costs for amorphous silicon, costs for a fully developed polycrystalline thin-film technology are likely to be dominated by the costs of a sheet of glass onto which the polycrystalline thin films are deposited and, perhaps, a second sheet of glass to complete an environmentally-protected PV sandwich. The thin films are polycrystalline, again unlike single-crystal silicon, and both CIS and CdTe devices have achieved 12% total area efficiencies while showing promise of stabilities similar to those of conventional crystalline silicon technology. Figure 2 shows the arrangement of atoms in the chalcopyrite crystal structure of CIS material (2). Finally, researchers have explored techniques for making efficient thin-film devices from polycrystalline silicon; the films are actually quite thick since they are about 50 times thicker than those used in thin-film technologies. Figure 3 shows the nature of polycrystalline silicon material (3).

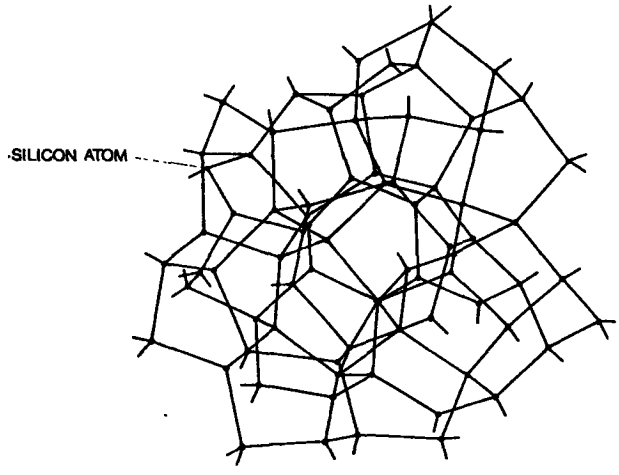


Figure 1. The lines connecting the dots (silicon atoms) represent the chemical bonds in amorphous silicon material

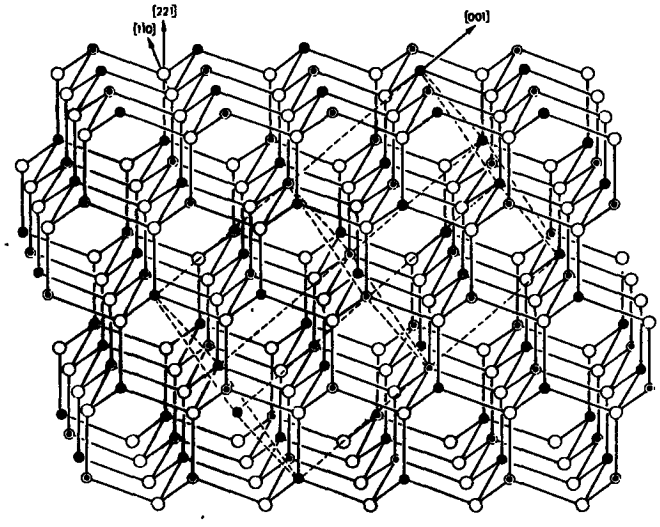


Figure 2. Chalcopyrite crystal lattice structure, (110) projection. Unit cell is indicated by dashed lines.●, Cu;⊙, In;○ , Se.

CRYSTALLINE TECHNOLOGIES

So-called III-V technologies are crystalline materials based on elements from columns III and V of the periodic table. Gallium arsenide (GaAs) is the most widely studied representative of these materials. III-V materials are highly efficient solar absorbers; thin films are sufficient for making PV devices. However, it is costly to make single-crystal thin-films using III-V materials, so III-V PV devices, which have had solar conversion efficiencies well above 30%, may be applied to solar concentrator technologies in which costly, small-area, high-efficiency devices can be tolerated. Indium phosphide (InP) is another III-V material for which high-efficiency devices have been made. InP may be important for PV applications in space because it is relatively resistant to damage from radiation. Figure 4 shows the orderly single-crystal arrangement of atoms in the zinc blende structure characteristic of III-V materials (4).

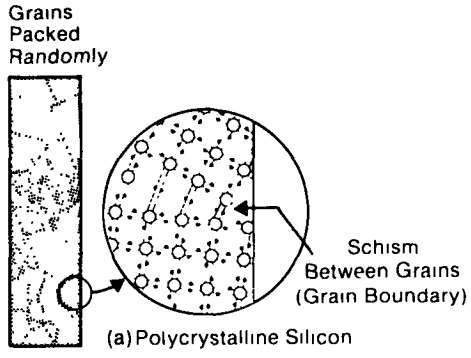


Figure 3. **Polycrystalline silicon is made up of randomly packed grains, each of which is a single crystal of silicon.**

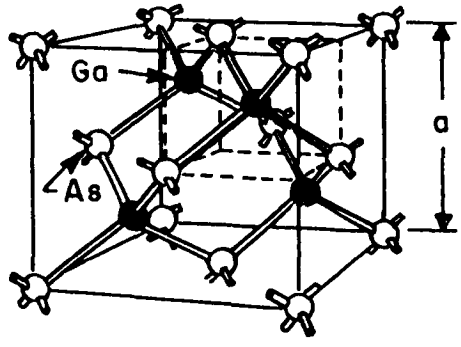


Figure 4. Zinc Blende (GaAs, GaP InSb, etc.) Single crystal structure.

PV DEVICES

Regardless of the technology or crystallinity of the materials, the heart of a PV device is a voltage arising between the junction of two or more layers having different electrical properties. For example, an n-type amorphous silicon layer in contact with an undoped intrinsic layer in contact with a p-type amorphous silicon layer results in a voltage within the layers that acts on electrons freed by the sunlight incident on the device. Electrical contacts over the entire back area of the device, and either narrow strip contacts or transparent electrical contacts over the front area, connect the PV device to loads or electric power conditioners. The junction in a CIS device comes from contact with another polycrystalline thin-film layer of CdS, whereas the junction in a GaAs device is often designed as that resulting from contact between n-type and p-type layers. A significant challenge in making PV devices is preparing these layers over large areas while maintaining homogeneity in the material properties of the layers. Figure 5 shows different junction possibilities for PV devices (3).

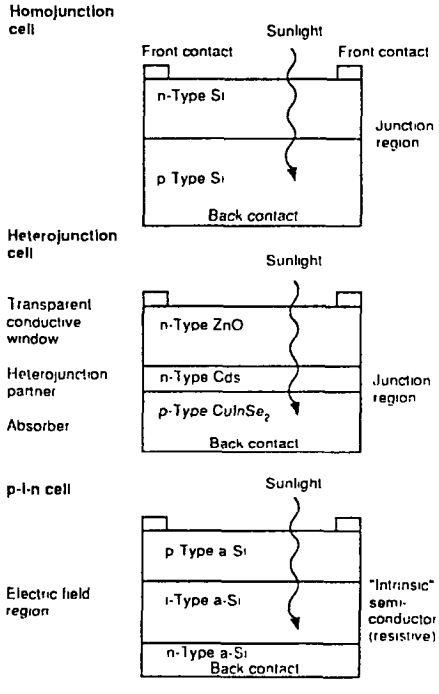


Figure 5. Different solar cell structures.

RESEARCH PROGRESS

Figure 6 shows the increase in conversion efficiency for different PV technologies over the past decade (6). It is possible that progress can continue for another decade at the same rate because the theoretical conversion efficiencies of PV devices are much higher than the efficiencies shown in Figure 6. Figure 7 shows the decrease in the cost of

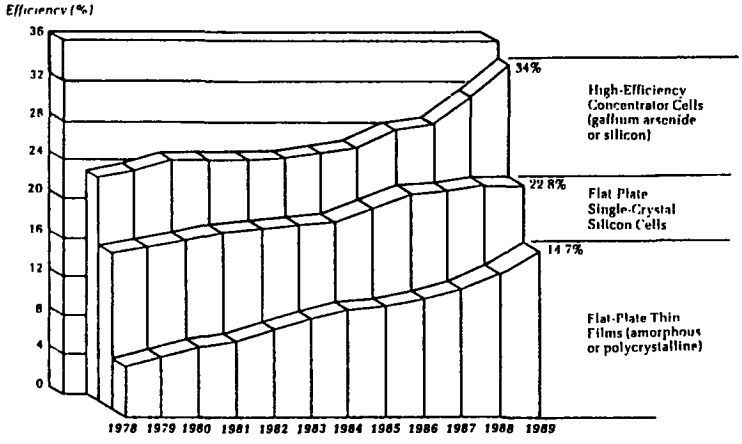


Figure 6. The efficiencies of laboratory cells increased markedly from 1978 to 1989.

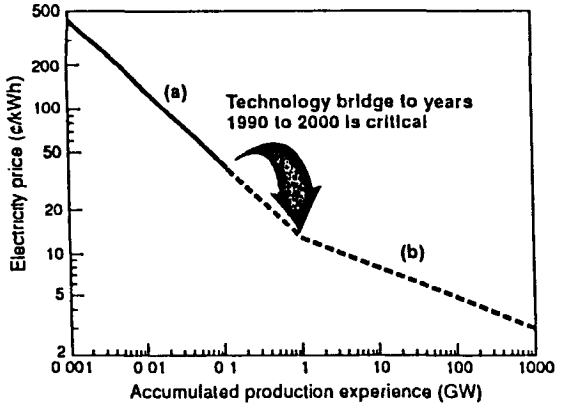


Figure 7. Dramatic cost reductions should continue through the end of the century as production increases. (a) price reduction of 68% achieved per tenfold increase in production experience; (b) price reduction of 40% anticipated per tenfold increase in production.

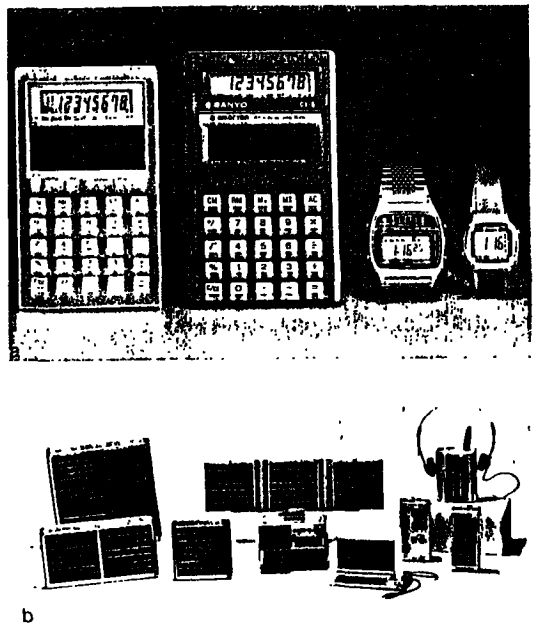


Figure 8. Example of consumer applications: (a) pocketable calculators and wrist watches, and (b) pocketable tape recorder, television, radio, and battery chargers. (Courtesy of Fuji Electric Co., Ltd.)

electricity produced by PV systems over the past decade and the expected future cost decreases that will occur as production levels increase (7). An early study suggested that these cost reductions were well within the potential of PV technologies, especially the thin-film technologies, because of their greatly reduced materials requirements (8). Efficiency increases and cost reductions, both past and future, provide the strong interest in PV by companies and countries around the world.

COMMERCIAL APPLICATIONS

Consumer products were a highly visible application for PV technologies during the 1980s. Solar-powered calculators, with annual sales of about 100 million calculators, are usually powered by amorphous silicon thin films (5). Solar-powered watches and clocks are among other consumer product applications (Figure 8) for amorphous

Figure 9. These outdoor lights are powered by photovoltaics.

Figure 10. PV-powered global positioning system at China lake.

silicon PV devices (9). Sailboat owners have bought PV power supplies for years, while car sunroofs, street address lighting, and outdoor lighting (Figure 9) are consumer products leading to new markets (3) in this area.

So-called stand-alone PV applications are those applications not connected to an electric utility. Military applications provide many good examples--Figure 10 shows a PV-powered global positioning system used by the Navy (10). The Navy performed a study in 1986 identifying more than 21,000 cost-effective PV applications (3). Military applications of PV for communications or navigation often incorporate diesel engine units for backup. Another stand-alone application--water pumping--is shown in Figure 11 (11). This stand-alone application of PV would not necessarily have diesel backup when there is water storage. Three additional noteworthy aspects of Figure 11 are: 1) the module is made from polycrystalline thin-film CdTe, 2) the manufacturer is British, and 3) the customer is international--Saudi Arabia, in this case. Telecommunications, village power for Third World villages, warning signals, remote monitoring, isolated lighting, cathodic power protection, and many other remote power requirements provide general categories for thousands of stand-alone applications of PV (12).

Figure 11. A 54 W thin-film CdTe array for water pumping deployed by BP Solar in Saudi Arabia.

Grid-connected or utility applications are an important category in which the long-term impact of PV can have enormous importance for the United States and the world because of the size of the market. Utility systems (Figure 12) have been relatively small, with only a handful of systems built on the order of a megawatt (MW) or so in size (3). Instead, utilities have identified high-value energy markets yielding thousands of smaller systems installed within utilities to supply microwave repeaters, cathodic protection of pipelines, telemetering, lighting, remote switching, etc. (6). The Electric Power Research Institute estimates

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

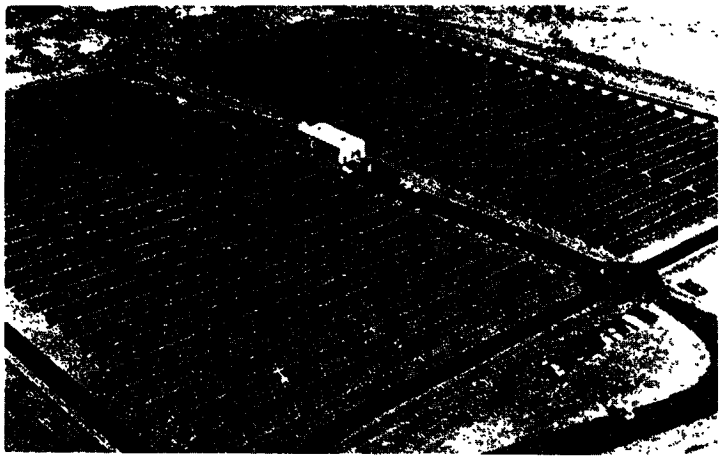


Figure 12. Near Sacramento, California, these arrays produce about a megawatt of power for the Sacramento Municipal Utility District.

that the potential utility-owned PV system market will be some 40,000 installations totalling over 11 MW by the year 1996 (13). Grid-connected houses (Figure 13), PV-powered pumps for swimming pools, and grid-connected commercial buildings are of interest for some utilities (14) because of land limitations or because of the demand-side management opportunities in these applications.

One of the oldest PV markets, although small, is for space system power supplies (Figure 14) because PV is very cost effective for satellite applications. A rapidly increasing market is the international market, because PV can provide village lighting, water pumping, and refrigeration for medical supplies more reliably and cheaper than small diesel or gasoline generators. This international market is not just an equatorial market, but rather one where remoteness and need combine to make PV a cost-effective choice (Figure 15) (15). Finally, spin-off applications of PV R&D include thin-film transistors of amorphous silicon and optoelectronic devices using III-V materials.

International market results are shown in Figure 16, which shows worldwide PV shipments and country shares of this market activity (7). Note that a recent estimate for the 1991 world market sales is between 58 and 62 MW (16). This steady market growth of PV has attracted companies throughout the world. The largest of all future markets is expected to be the U.S. utility markets. Although market forecasts are highly uncertain, one prediction for the U.S. utility market alone is between 1 GW and 5 GW installed during the next decade, with almost 500 GW cumulative installed by the year 2030 (12, 17). This market size is big enough to interest most large companies, because it represents a $1 trillion market over the 40-year period.

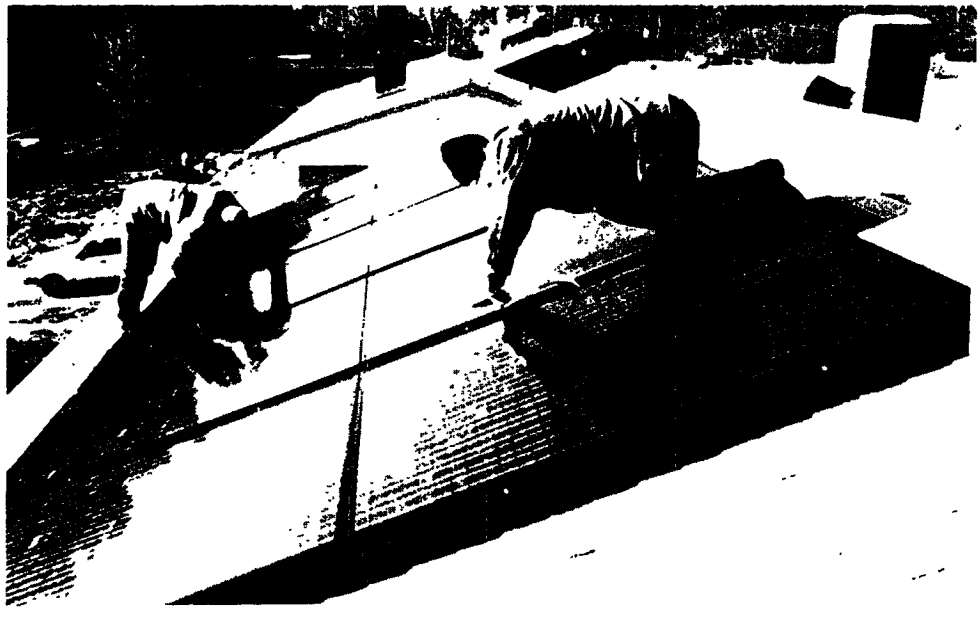


Figure 13. The Gardner study has earned a high degree of confidence in PV's system reliability and interaction with the distribution system.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

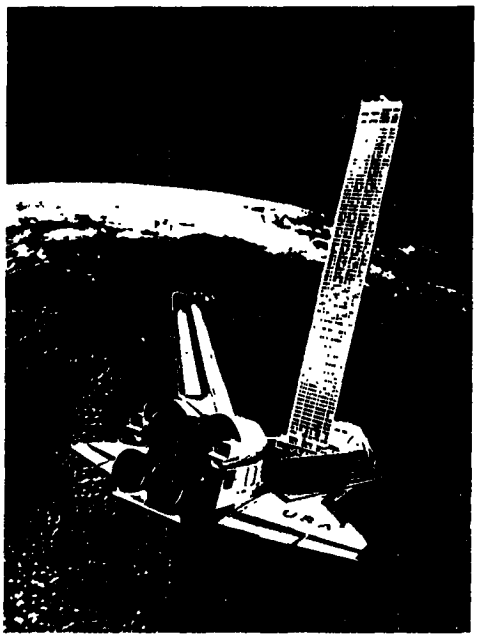


Figure 14. Discovery Shuttle with Lockheed solar panel.

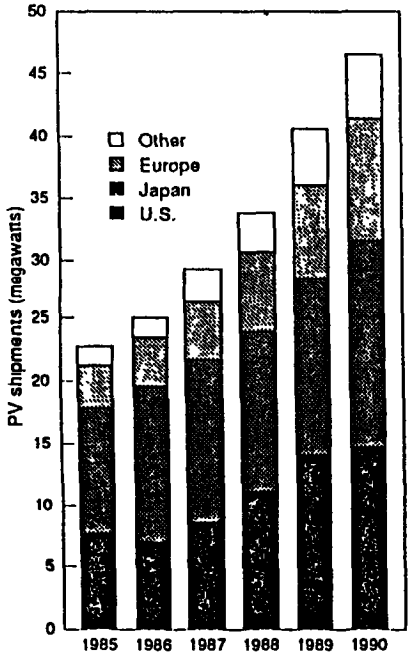


Figure 16. The U.S. share of the photovoltaics market has climbed to about 35%. (Source: PV News, February 1991)

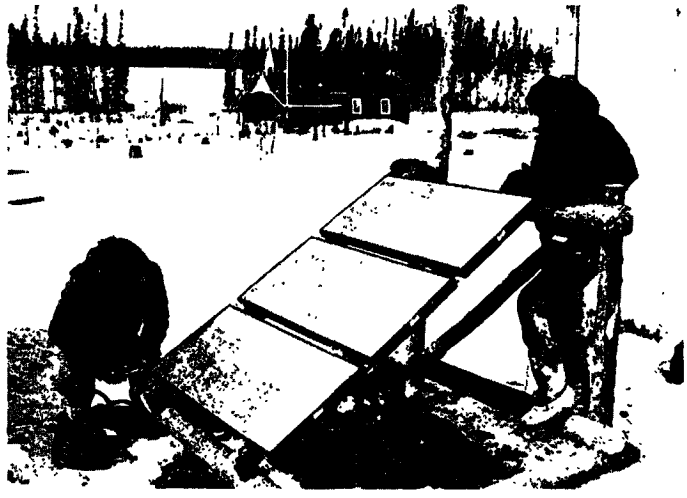


Figure 15. A photovoltaic power system in a remote community of Canada's far north

TECHNOLOGY TRANSFER

The transfer of a technology developed with federal funds can take place in several ways. One example would be a U.S. industry licensing a technology developed by NREL. Another example could come from cooperative R&D projects involving NREL and industry researchers. As federal laboratories have tried to help U.S. industry in the international marketplace, seven technology transfer mechanisms have come into use. They are: 1) subcontracted R&D to industry using federal/NREL funds, 2) cooperative R&D agreements (CRADAs), 3) industry-sponsored R&D (what NREL calls Work for Others), 4) user facilities within NREL, 5) technology licenses, 6) research exchanges, and 7) information dissemination through research publications, workshops, and conferences. For over a decade, as mentioned earlier, NREL has worked closely with U.S. industry to develop PV technologies through R&D subcontracts. The most recent mechanism, CRADAs, arose from the National Competitiveness Technology Transfer Act of 1989 (18). A common CRADA might involve NREL scientists working with private industry scientists on an agreed-upon research project involving no exchange of funds. A signed agreement provides protection for intellectual property resulting from the research project for 5 years; i.e., the research information is protected from Freedom of Information Act inquiries. CRADAs are becoming today's currency for technology transfer in the national laboratories, consistent with the following definition for technology transfer: "Technology transfer is collaborative research and development between laboratory researchers and industry researchers for the purpose of aiding industry's commercialization of products and services." CRADAs can involve more than one private sector entity so that joint ventures

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

or consortia of national laboratories, PV manufacturers, electric utility suppliers, and electric utility end users are possible. Such vertically integrated joint ventures or consortia have the potential to be the government/private sector partnerships of the future.

As new photovoltaic technologies come of age and NREL researchers reach out to U.S. industry, NREL sees a period of even closer interaction with the U.S. private sector for the purpose of creating the nation's photovoltaic future.

REFERENCES

1) Yoshihiro Hamakawa, "Photovoltaic Power," Scientific American, Vol. 256, No. 4, pp. 86-92 (April 1987).

2) Lawrence L. Kazmerski and Sigurd Wagner, "Cu-Ternary Chalcopyrite Solar Cells," Current Topics in Photovoltaic, Edited by T. J. Coutts and J. D. Meakin, pp. 41-109 Academic Press (1985).

3) Solar Energy Research Institute, "Photovoltaic Technical Information Guide," Second Edition, SERI/SP-320-3328 (1988).

4) Stephen J. Fonash, Solar Cell Device Physics, p. 8, Academic Press (1981).

5) D. E. Carlson, "Overview of Amorphous Silicon Photovoltaic Module Development," Solar Cells, Vol. 30, pp. 277-283 (1991).

6) John P. Thornton, Richard DeBlasio, and Kenneth Zweibel, "Photovoltaic--Today's Reality, Tomorrow's Promise," Energy Engineering, Vol. 87, No. 3, p. 62 (1990).

7) U.S. Department of Energy, "Photovoltaic Program Plan", DOE/CH10093-92 (to be published).

8) J. Stone, E. Witt, R. McConnell, T. Flaim, T. Surek, and
D. Ritchie, "Recent Developments in Photovoltaic," Proceedings of the Seventeenth IEEE Photovoltaic Specialists Conference, Kissimmee, FL, pp. 1178-1183 (May 1984).

9) Yoshihiro Hamakawa, "Amorphous-Silicon Solar Cell," Current Topics in Photovoltaic, Edited by T. J. Coutts and J. D. Meakin, pp. 111-168, Academic Press (1985).

10) Roch Ducey, "Recent Activities of the U.S. Department of Defense Photovoltaic Review Committee," Proceedings of the Biennial Congress of the International Solar Energy Society, Denver, CO, Vol. 1, Part 1, pp. 271-276 (August 1991).

11) Harin S. Ullal, Kenneth Zweibel, Richard L. Mitchell,
Rommel Noufi, "Polycrystalline Thin Film Photovoltaic Technology," Proceedings of the Biennial Congress of the International Solar Energy Society, Denver, CO, Vol. 1, Part 1, pp. 3-8 (August 1991).

12) Ken Zweibel, Harnessing Solar Power: The Photovoltaic Challenge, Plenum Press (1990).

13) "On-Site Utility Applications for Photovoltaic," EPRI Journal, Vol. 16, No. 2, pp. 26-37 (March 1991).

14) John J. Bzura, "Residential Photovoltaic: The New England Experience Builds Confidence in PV," Photovoltaic: New Opportunities for Utilities, pp. 2-5, DOE/CH10093-113 (July 1991).

15) Per Drewes, "Capturing the Sun," Proceedings of the Biennial Congress of the International Solar Energy Society, Denver, CO, Vol. 1, Part 1, pp. 525-530 (August 1991).

16) Photovoltaic Insiders Report, Vol. X, No. 8, p. 6 (August 1991).

17) Idaho National Engineering Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Sandia National Laboratories and Solar Energy Research Institute, "The Potential of Renewable Energy: An Interlaboratory White Paper," SERI/TP-260-3674, Appendix G (March 1990).

18) Public Law 101-189, PART C (1989).

SOFTWARE ENGINEERING

(Session D6/Room C4)

Thursday December 5, 1991

- **Software Reengineering**
- **COSTMODL: An Automated Software Development Cost Estimation Tool**
- **Increasing Productivity Through Total Reuse Management**
- **How Hypermedia Can Increase the Productivity of Software Development Teams**

PRECEDING PAGE BLANK NOT FILMED

N92-22708

SOFTWARE REENGINEERING

Ernest M. Fridge III
Deputy Chief, Software Technology Branch/PT4
NASA/Johnson Space Center
Houston, Texas 77058
(713) 483-8109
Nasamail: EFRIDGE

ABSTRACT

Today's software systems generally use obsolete technology, are not integrated properly with other software systems, and are difficult and costly to maintain. The discipline of reverse engineering is becoming prominent as organizations try to move their systems up to more modern and maintainable technology in a cost effective manner. The Johnson Space Center created a significant set of tools to develpop and maintain FORTRAN and C code during development of the space shuttle. This tool set forms the basis for an integrated environment to reengineer existing code into modern software engineering structures which are then easier and less costly to maintain and which allow a fairly straightforward translation into other target languages. The environment will support these structures and practices even in areas where the language definition and compilers do not enforce good software engineering. The knowledge and data captured using the reverse engineering tools is passed to standard forward engineering tools to redesign or perform major upgrades to software systems in a much more cost effective manner than using older technologies. A beta version of the environment was released in March, 1991. The commercial potential for such reengineering tools is very great. CASE TRENDS magazine reported it to be the primary concern of over four hundred of the top MIS executives.

INTRODUCTION

Programs in use today generally have all of the functional and information processing capabilities required to do their specified job. However, older programs usually use obsolete technology, are not integrated properly with other programs, and are difficult to maintain. Reengineering is becoming a prominent discipline as organizations try to move their systems to more modern and maintainable technologies. Johnson Space Center's (JSC) Software Technology Branch (STB) is researching and developing a system to support reengineering older FORTRAN programs into more maintainable forms that can also be more readily translated to a modern language such as FORTRAN 8x, Ada, or C. This activity has led to the development of maintenance strategies for design recovery and reengineering. These strategies include a set of standards, methodologies, and the concepts for a software environment to support design recovery and reengineering.

This document provides a brief description of the problem being addressed and the approach that is being taken by the STB toward providing an economic solution to the problem. A statement of the maintenance problems,the benefits and drawbacks of three alternative solutions, and a brief history of the STB's experience in software reengineering are followed by the STB's new FORTRAN standards, methodology, and the concepts for a software environment.

STATEMENT OF THE PROBLEM

Based on trends in the computer industry over the last few years, it is clear that computer hardware, languages, and procedures are not static. The software industry recognizes that a large existing software base must be dealt with as new software engineering concepts and software technologies emerge. The old systems use outdated technology and are costly to maintain. At JSC, as in industry at large, there is a large investment in existing FORTRAN software. These FORTRAN systems do not consistently use modern

 PRECEDING PAGE BLANK NOT FILMED

software practices that can increase maintainability. Yet these systems must be maintained for perhaps the next 20 years. Management is seeking ways to reduce maintenance costs.

In the 1960s-70s many FORTRAN programs were developed at JSC, each with its own sizeable software development team, and its own input/output format. These programs could not communicate readily and eventually were "wired" together in a very crude semblance of integration. Standards could not be enforced because FORTRAN did not enforce them and some were not visible by just looking at the code. The problem was aggravated by the lack of training of new developers plus a 50 percent turnover in the very large development staff every two years. In addition, the user organizations had more people doing development than the development group, and these other organizations were not always aware of the standards and support tools available. This history has left JSC with the following problems:

- Many programs are large and difficult to understand, resulting in maintenance problems.

- The problems in maintenance led to users keeping their own versions of programs, resulting in tremendous duplication.

Many of the FORTRAN programs have already been converted from their original dialect of FORTRAN to the FORTRAN 77 standard. Additional conversions will periodically be required even if only to new FORTRAN standards. It is necessary to consider the question, where will that code have to be in five or ten years? Three possible answers come to mind:

- FORTRAN 77 is the current standard, but this will be replaced by newer Fortran standards. As vendors stop supporting FORTRAN 77, existing FORTRAN will have to move to the new standard or to another language.

- Much of the code may move to the Ada language. This will be particularly true on Space Station Freedom work.

- With C being the language of choice for Unix , some of the code might move to the C language.

ALTERNATIVE SOLUTIONS

Three alternative solutions to the problems identified above have been identified: complete redevelopment of the program, code translation to a more modern language or version of a language, and reengineering. Each of these is illustrated in figure 1 and discussed briefly in the following paragraphs.

Redevelopment of a system from scratch is very expensive. Redevelopment includes all of the same phases of the life cycle as new development, from requirements through integration and testing. Extensive domain analysis is required, and there is a risk of incomplete requirements. All too often it is reported that a large program will be redeveloped from scratch to a more modern style only to find out that the new developers did not understand all of the functions and necessary information requirements of the existing system.

Code translation, especially automatic code translation, costs much less. Some might then ask, why worry about all of this now? We can use a translator when the time comes that we are forced to move the code for-

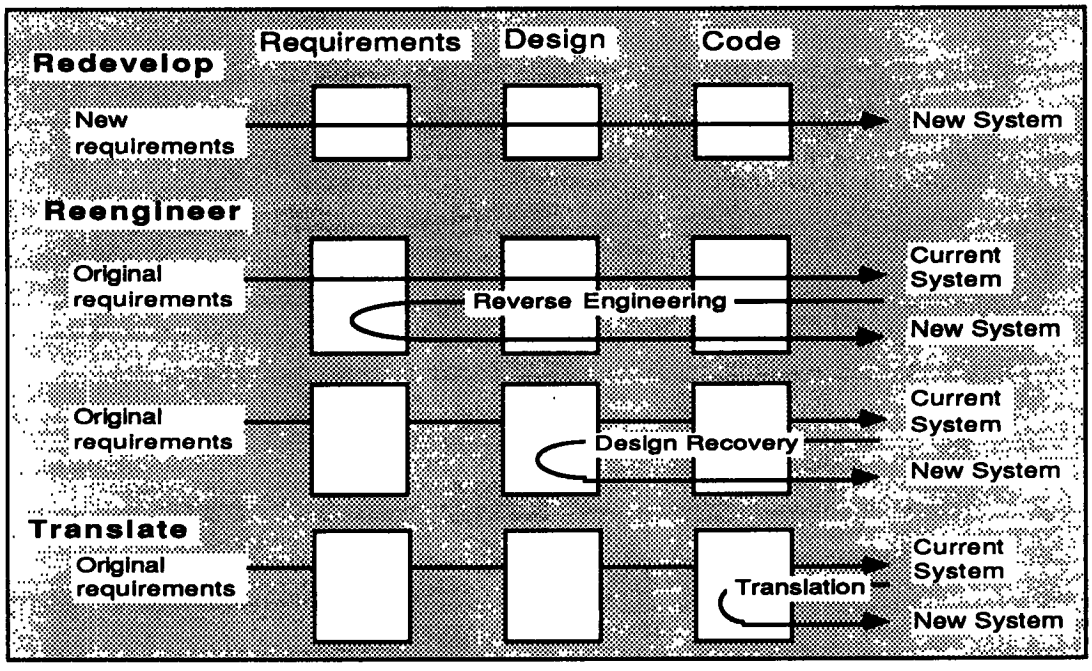


Figure 1. Alternative Solutions

ward. Although this would be a nice solution, the truth is that code translators have proven unsuccessful due to several major reasons:

- Poor existing control flow is translated into poor control flow.
- Poor existing data structures remain poor data structures.
- Input/output translation usually produces hard to read "unnatural"code in the new language.
- Translation does not take advantage of the code and data packaging techniques available in the newer languages. Attempts to automatically translate some FORTRAN programs to Ada have failed.

Reengineering is the combination of "reverse engineering" a working software system and then "forward engineering" a new system based on the results of the reverse engineering. Forward engineering is the standard process of generating software from "scratch." It is composed of the life cycle phases such as requirements, architectural design, detailed design, code development, testing, etc. In each phase, certain products are required and the activities which produce them are defined. Each product is required to be complete and consistent. To progress forward to a new phase normally requires a new representation of the products which involve more detail such as new derived requirements, design decisions, trade off evaluation between alternative approaches, etc. Finally, code is developed which is the most complete, consistent, and detailed representation of the required product.

Reverse engineering is the reverse of forward engineering. It is the process of starting with existing code and going backward through the software development life cycle. Life cycle products are, therefore, obtained by abstracting from more detailed representations to more abstract ones. This process should proceed much faster than forward engineering since all of the details required are available. Reverse engineering starts with the most detailed representation, which has also proven to be complete and consistent since it can currently do the job required. Developing products in reverse involves abstracting out only the essential information and hiding the non-essential details at each reverse step.

How far to go backward in the reverse engineering process before it is stopped and forward engineering begins is a critical question and involves trade offs. It is important to understand all of what the program

does, all of the information it handles, and the control flow since these are probably required to get the job done. This implies taking the reverse process far enough to understand *what* the "as is" program is. This is usually more significant than *how* the program does its job since the *how* is usually the part that will be changed in any following forward engineering process.

What a program does is called its *requirements*. *How* it meets those requirements is its *design*. For a reverse engineered program it is the design that will be updated more often than *what* the program will do. Modern software engineering techniques and technologies such as user interfaces, database management, memory utilization, data structuring, packages, objects, etc. will affect the design, not *what* the program does. Therefore, once it is understood what the program does and what is obsolete, then the forward engineering process can begin with confidence.

Reverse engineering is referred to as "design recovery" when the reverse engineering process stops at the recovery of the design of the implementation, rather than proceeding on to a higher level of abstraction to include the recovery of the requirements. The basic process of this level of design recovery involves recovery of information about the code modules and the data structures in an existing program. This information will support the programmer/analyst who is maintaining an unfamiliar large FORTRAN program, upgrading it for maintainability, or converting it to another target language.

However, a better job of redesigning a program can be accomplished with requirements recovery than with design recovery. To carry the reverse engineering process beyond design recovery to requirements recovery is difficult and requires higher levels of domain knowledge to do the abstractions. The *whys* of the requirements, design, and implementation can only be provided by someone very familiar with the program and the domain. This level of expertise is often very difficult to find and have dedicated to the reengineering process. For this reason, the methods and tools that the STB has developed initially assume reverse engineering only to the design recovery stage. Future development will be based on feedback from the JSC software engineering community. The current standards, methods, tools, and environment are all designed to be sufficiently flexible and extendible to enable the strategies to be extended to cover the full spectrum of reverse engineering.

The overriding philosophy of this planned reverse engineering process is to capture the total software implementation in an electronic form. This includes source code, documentation, databases, etc. Figure 2 illustrates the progression of data structures from COMGEN-compatible code (see section "Software Technology Branch's Reengineering History") to reengineered code. This progression in electronic form ensures that the total consistent and complete requirements representation is available. Software tools are provided to support the generation of the more abstract products required for engineering in reverse as well as capturing rationale and decisions of the engineer. By the continuing process of abstracting the information about the program into the different representations, the engineer can remain more confident that information is not being lost or inadvertently "falling through the cracks."

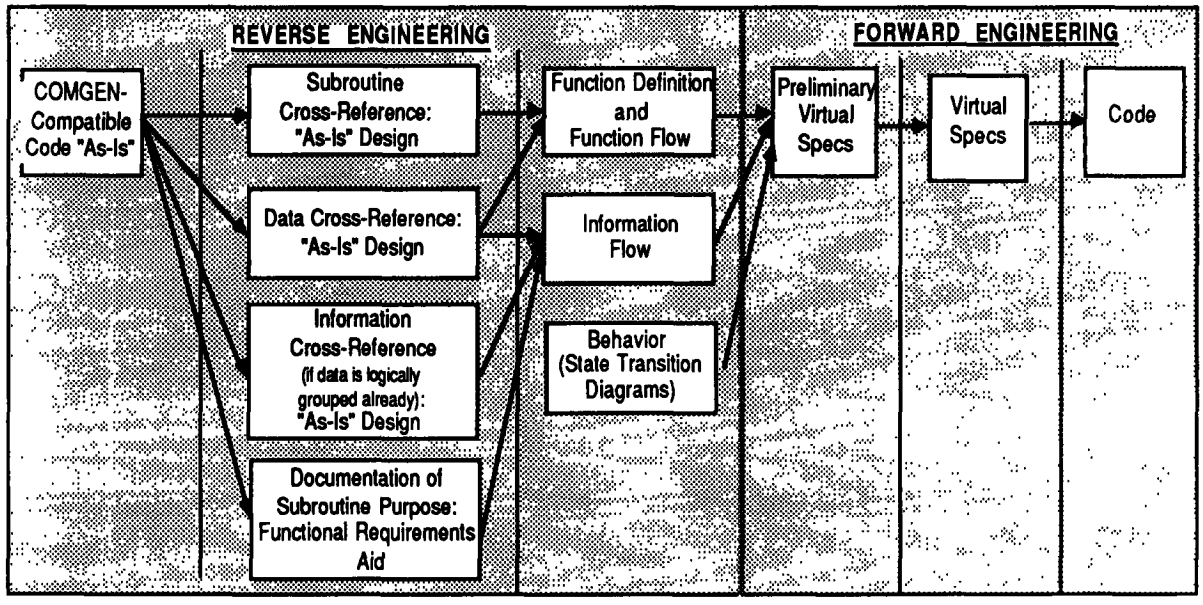


Figure 2. Data Structure Progression

SOFTWARE TECHNOLOGY BRANCH'S REENGINEERING HISTORY

In the early 1970's, the Mission Planning and Analysis Division's (MPAD) Software Development Branch and TRW/Houston developed a tool, called COMGEN, that began as a COMMON block specification statement generator. It grew to include many other functions as new techniques were developed. Later COMGEN was broken up into a continually evolving set of tools with common data interface structures. This tool set supports the maintenance of FORTRAN programs today on Unisys and multiple Unix systems. People still refer to this tool set as COMGEN tools, and a program that complies with the MPAD standard COMMON concept as a COMGEN-compatible program. [1,2,3]

In the 1970's, MPAD performed a lot of software reengineering to meet the goal of combining many of the independently developed engineering programs, each with its own input/output formats. Many of the modern concepts such as separation of input/output processing from the applications, databases, data structures, packages, generics, objects, etc. were recognized and simulated to some degree. They were not called by the modern names, of course, but the design engineers were trying to do good engineering, modularization, and data handling. Even though these techniques were known in the 1970's, they are just now really becoming popular because of newer technologies such as database management systems, user interface tools sets, and modern languages that actually embed and enforce good software engineering practices.

In the late 1980's, some of the personnel and the functions of the Software Development Branch were reorganized into the newly created Software Technology Branch (STB). The STB's reengineering history has put JSC in a better position with respect to the maintainability of its older software than many other organizations. The positive results of this experience include the following:

- Most of the software is reasonably modular.
- The data has some structure.
- Most of the software at JSC is reasonably compatible with the STB's tools, including the in-line documentation.
- The large complex programs that support many simulations have considerable software reuse and information sharing.

MAINTENANCE STRATEGIES

The strategies presented in this document are intended to help with design recovery in support of programmer/analysts who are required to maintain large FORTRAN programs that they did not develop. In addition, these strategies are intended to support reengineering of existing FORTRAN code into modern software engineering structures, which are then easier to maintain and which allow a fairly straight forward translation into other target languages. The STB is proposing standards, methods, and an integrated software environment based upon the significant set of tools built to develop and maintain FORTRAN code for the Space Shuttle. [4,5,6,7,8] The environment will support these structures and practices even in areas where the language definition and compilers do not enforce good software engineering practices.

New FORTRAN Standards

New standards, which allow modern software engineering constructs to be used in FORTRAN 77, have been defined by the STB. [5] These standards are added to existing standards defined by the former MPAD and still in use in the mission planning and analysis domain. The goal of the new standards is to improve maintainability and permit relatively automated translations to newer languages. In table 1, the standards and their benefits are summarized. These standards address documentation, longer variable names, modern control flow structures, grouping subprograms together as virtual packages, data structuring, and input/output encapsulation in separate subprograms. Where FORTRAN 77 does not provide the constructs, virtual constructs are provided along with a tool environment to support their development and maintenance. The existing core of FORTRAN programmers should have little problem with the standards and new FORTRAN code should adhere to them from the start.

Table 1. Standards Summary

Standard	Benefit
Documentation	
Header statement before code blocks	Understandability
Requirements in CD1 statements	Understandability and traceability
Rationale in CD7 statements	Design knowledge capture
Virtual package identification	Maintenance
Longer, more meaningful variable names	Understandability
Modern control flow structures	Maintenance and understandability
Block DO	
DO WHILE	
Grouping subprograms into virtual packages	Higher level of abstraction, understandability
Data structuring	
Preferred use of calling parameters	Maintenance
Controlled use of COMMON blocks	Maintenance
	INCLUDE
	COMMON database concept
Preferably encapsulate input/output in separate subprograms	Maintenance and support to future conversions

Design Recovery and Reengineering Methodology

The reengineering methodology defines the steps, the skills required, and guidelines on how far to reverse engineer before deciding to rebuild. The key goal is to update to modern technology and software engineering concepts without losing required functions and data. Methods are provided that have the flexibility to meet multiple levels of conversion, each of which improves maintainability. Figure 3

illustrates five methods. [6] Method 1 converts an arbitrary FORTRAN program to COMGEN-compatible FORTRAN, which provides in-line documentation, data structure, and unique data names within a COMMON structure. Method 2 converts software already in this format to the new "standard" FORTRAN with a more Ada-like structure that is ready for a mostly automated translation by Method 3 to a target language that embeds software engineering principles. Alternatively, COMGEN-compatible programs can be converted directly to a target language like Ada by Method 4. Although it is easier to convert a FORTRAN program when the code already meets the standard COMMON concept, commonly known as COMGEN-compatible, arbitrary FORTRAN can be directly converted to a target language by Method 5.

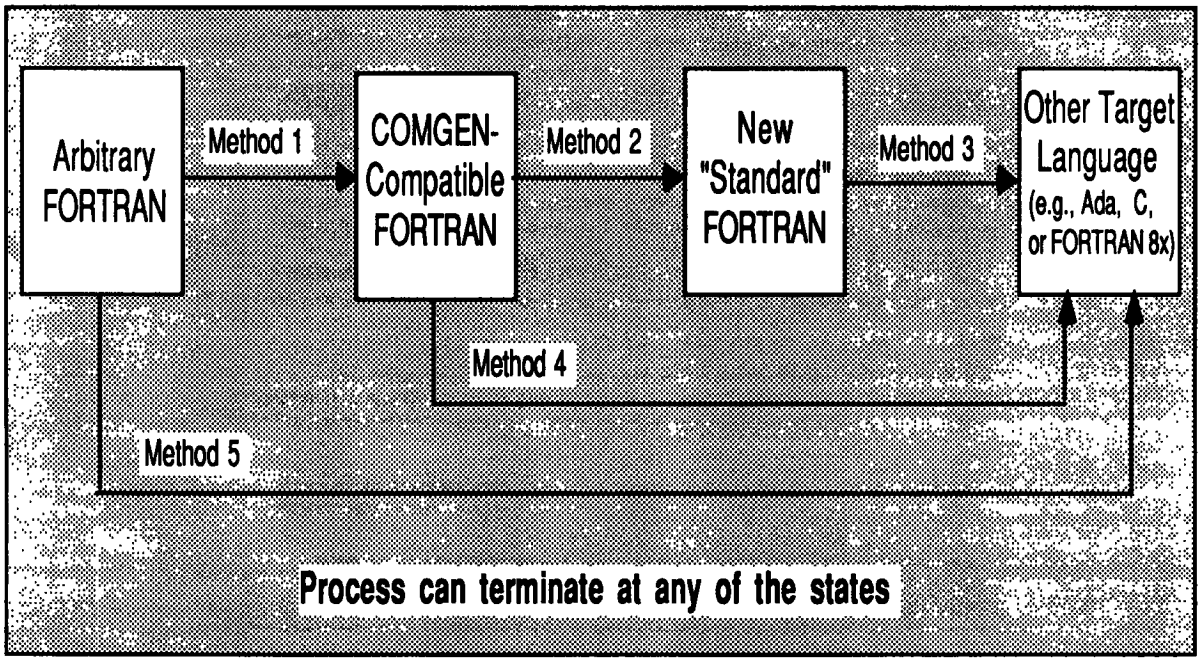


Figure 3. Reengineering Methods

Environment to Support Design Recovery and Reengineering

The STB's reengineering environment [7] is being built around three components: standards, methods, and tools that support the standards and the methods. It contains modified versions of the tools used to support the current JSC FORTRAN programs plus commercial off-the-shelf (COTS) tools and additional custom-built tools. The intent is to get an environment out into use in JSC's maintenance community to provide support for upgrading FORTRAN programs in terms of maintainability in the near-term, then to extend the functionality of the tool set and environment in response to feedback from the programmers/ analysts. Currently about eight groups at JSC are using the tools. Some support for the C language exists and a cooperative agreement with the Microelectronic and Computer Technology Corporation (MCC) is evaluating research into design recovery of C programs.

The environment has been designed with stable interfaces defined to provide for the maximum degree of seamlessness that is desirable. It is doubtful that COTS tools can be integrated seamlessly into the environment as no standard interfaces have yet been established for either user interface or data interface (as opposed to data exchange). The tools are integrated at the front end by a user interface and behind the screen by two logical databases, one containing data passed to and from the tools and the other containing the original and modified source code as shown in figure 4. CASE framework tools are being evaluated as possible integration mechanisms.

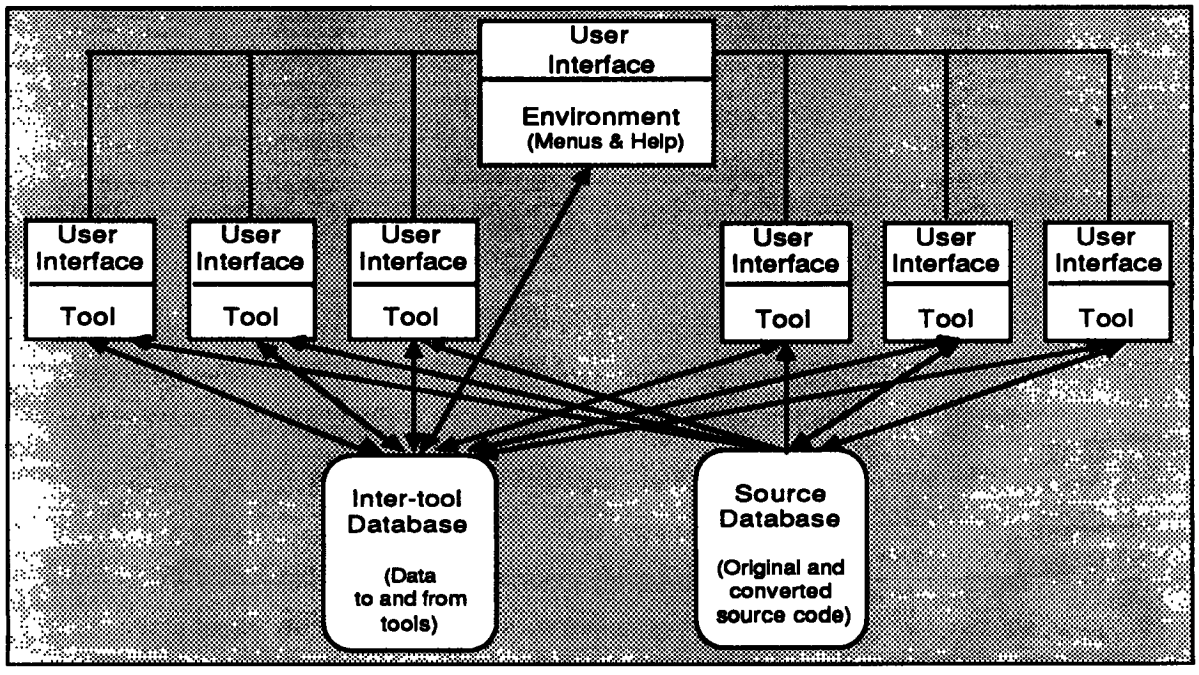


Figure 4. Conceptual Architecture of the Design Recovery and Reengineering Environment

The environment will not be a completely automated environment since much work will still have to be done by a programmer/analyst. A person must be in the loop to provide the required puzzle-solving skills that are beyond the capabilities of state-of-the-practice tools. However, as an experience base is accrued in design recovery and reengineering, knowledge-based capabilities can be added to the environment.

Version 1 of the environment called REengineering APplications (REAP) was delivered in June, 1991. This integrated all existing JSC supported tools listed above, behind a common user interface built on the MOTIF standard. It contains major elements of all subsystems and encapsulates the capabilities that have been developed and used at JSC during the last fifteen years. A version with improved tool integration, user interface enhancements, and the commercial LOGISCOPE tool was delivered in October, 1991. The Fortran design recovery version should be available in February, 1992. MCC should also have delivered an evaluation prototype of a design recovery capability for the C language by that time. In parallel, the study of using CASE framework standards and tools to better integrate and manage this environment should be completed early in 1992 and the version 2 series will be delivered on one of these platforms. The plans and design of REAP are such, that all deliveries containing COTS products will be tailorable so that users can delete the COTS tools that they do not want to liscense. This policy even includes the framework integration tools. In most cases, similar functions might still be available but they would have less capability.

CONCLUSIONS

JSC has a large amount of exsting code in FORTRAN that embodies domain knowledge and required functionality. This code must be maintained and eventually translated to more modern languages. Three primary alternative solutions have been identified to address the maintenance problems of these old FORTRAN programs: complete redevelopment of the programs, code translation to a more modern language or version of a language, and reengineering. Complete redevelopment is effective but very costly. Simple code translation is cheap, but usually ineffective since seldom do the old systems incorporate modern software engineering concepts such as good data structuring, good control structuring, packages, objects, etc., that should be present in the new system. Modern languages such as Ada have constructs for representing these features, but translators cannot determine these features in the original code to map them into the new system. Reengineering is being recognized as a viable option because the old systems, in

spite of obsolete technology, do contain all of the required functionality and can get the job done. However, at the present time there are only a few expensive Computer Aided Software Engineering (CASE) tools and no total system environment available in the COTS market to support reengineering FORTRAN programs.

The STB maintenance strategies provide standards, methods, and a tool environment for upgrading current FORTRAN systems without losing the embedded engineering knowledge and at a lower cost than for complete redevelopment of the program. A useful environment for reengineering FORTRAN software can be built fairly quickly by building upon the existing FORTRAN development and maintenance tools, COTS products, new software and hardware technologies, plus current research into reuse, design recovery, and reengineering. This environment will support reengineering existing FORTRAN code into more maintainable forms that can also be readily translated into a modern language including newer versions of FORTRAN.

Two versions of the environment were delivered in 1991 which integrate the existing JSC tools plus the commercial LOGISCOPE tool behind a common MOTIF user interface. A Fortran design recovery capability should be available in February, 1992 and the MCC should deliver a design recovery prototype for the evaluation of design recovery in the C language by that time. Plans are to integrate this capability on a CASE framework tool during 1992.

GLOSSARY

arbitrary FORTRAN	FORTRAN program that is not compatible with the COMGEN standards long in place for JSC's mission planning and analysis domain.
COMGEN-compatible	FORTRAN program that is compatible with the COMGEN standards long in place for JSC's mission planning and analysis domain. [1]
COTS	Commercial-Off-The-Shelf
design recovery	Reverse engineering, the first step for maintenance or reengineering.
environment	Instantiation of a framework, i.e., an integrated collection of tools. It may support one or more methodologies and may also provide a framework for third party tools.
framework	Software system to integrate both the data and the control of new and existing tools; usual components include a user interface, object management system, and a tool set.
FORTRAN 77	ANSI standards for FORTRAN in effect in June 1990.
FORTRAN 8x	Future ANSI standards for FORTRAN; expected to be approved and released soon; draft standards have been circulated; unofficially called FORTRAN 90.
forward engineering	Process of developing software from "scratch," through the phases of requirements, design, and coding.
package	"A collection of logically related entities or computational resources" (Booch[9]).
reengineering	"The examination and alteration of a subject system to reconstitute it in a new form and the subsequent implementation of the new form" (Chikofsky and Cross [10]); combination of reverse engineering and forward engineering.
reverse engineering	"The process of analyzing a subject system to identify the system's components and their interrelationships and create representations of the system in another form or at a higher level of abstraction" (Chikofsky and Cross [10]); the first step of maintenance or reengineering; reverse of forward engineering; process of

	starting with existing code and going backward through the software development life cycle.
software maintenance	Process of modifying existing operational software while leaving its primary functions intact (Boehm [11]).
subject program	Program that is being maintained or reengineered.
virtual package	Package concept as defined by Booch [9], but implemented either in Ada, which enforces the concept, or in a language in which the concept must be supported procedurally.

REFERENCES

[1] Braley, Dennis: *Computer Program Development and Maintenance Techniques.* NASA IN 80-FM-55, NASA Johnson Space Center (Houston, TX), November 1980.

[2] Braley, Dennis: *Automated Software Documentation Techniques.* NASA Johnson Space Center (Houston, TX), April 1986.

[3] Braley, Dennis: *Software Development and Maintenance Aids Catalog.* NASA IN 86-FM-27, NASA Johnson Space Center (Houston, TX), October 1986.

[4] Fridge III, Ernest: *Maintenance Strategies for Design Recovery and Reengineering: Executive Summary and Problem Statement.* Volume 1. NASA Johnson Space Center (Houston, TX), June 1990.

[5] Braley, Dennis: *Maintenance Strategies for Design Recovery and Reengineering: FORTRAN Standards.* Volume 2. NASA Johnson Space Center (Houston, TX), June 1990.

[6] Braley, Dennis; and Plumb, Allan: *Maintenance Strategies for Design Recovery and Reengineering: Methods.* Volume 3. NASA Johnson Space Center (Houston, TX), June 1990.

[7] Braley, Dennis; and Plumb, Allan: *Maintenance Strategies for Design Recovery and Reengineering: Concepts for an Environment.* Volume 4. NASA Johnson Space Center (Houston, TX), June 1990.

[8] George, Vivian; and Plumb, Allan: *A Method for Conversion of FORTRAN Programs.* Barrios Technology, Inc. (Houston, TX), March 1990.

[9] Booch, G.: *Software Engineering with Ada.* Benjamin/Cummings Publishing Co., Inc. (Menlo Park, CA), 1983.

[10] Chikofsky, E. J.; and Cross II, J. H.: "Reverse Engineering and Design Recovery: A Taxonomy." *IEEE Software*, January 1990.

[11] Boehm, B. W.: *Software Engineering Economics.* Prentice-Hall (Englewood Cliffs, NJ), 1981.

U.S. Gov't

N92-22709

COSTMODL: AN AUTOMATED SOFTWARE DEVELOPMENT COST ESTIMATION TOOL

George B. Roush
NASA Johnson Space Center
Houston TX 77058

ABSTRACT

The cost of developing computer software continues to consume an increasing portion of many organizations' total budgets, both in the public and private sector. As this trend develops, the capability to produce reliable estimates of the effort and schedule required to develop a candidate software product takes on increasing importance. The COSTMODL program was developed to provide an in-house capability to perform development cost estimates for NASA software projects. COSTMODL is an automated software development cost estimation tool which incorporates five cost estimation algorithms including the latest models for the Ada language and incrementally developed products. The principal characteristic which sets COSTMODL apart from other software cost estimation programs is its capacity to be completely customized to a particular environment. The estimating equations can be recalibrated to reflect the programmer productivity characteristics demonstrated by the user's organization, and the set of significant factors which effect software development costs can be customized to reflect any unique properties of the user's development environment. Careful use of a capability such as COSTMODL can significantly reduce the risk of cost overruns and failed projects.

INTRODUCTION

Need for Formalized Software Cost Estimation

In the early days of the evolution of computer software, managers were forced to rely on the opinions of software development team leaders for estimates of the time and effort required to develop a candidate software product. This was a highly subjective process which was greatly influenced by factors such as the personality of the estimator (optimist or pessimist), pressure to underestimate to win a contract, etc. Since these estimates were based primarily on intuition and personal experience, the estimates were extremely difficult to reproduce or to refine as a project matured.

In the late 1970s, mathematical models began to emerge which attempted to quantify the parameters upon which an estimate was based, and to compute the estimates using equations which were developed based on actual data derived from completed software projects. The successful development of an estimating algorithm is dependent on the identification of the factors which affect productivity, and the careful collection of data over a sufficient period of time and number of projects to permit a valid statistical analysis. It is not surprising that the early models were hampered by the scarcity of good data.

Different Methodologies

As managers began to search for ways to improve the fidelity of their estimates, several methodologies emerged. Examples are Expert Opinion, Analogy, and Parametric Models. The expert opinion approach involves convening a group of experienced project managers to assess the known requirements for the new project, and to apply the knowledge resulting from their previous experience to collectively arrive at an estimate of the effort and schedule required to build the candidate product. This approach helps to minimize the effects of individual personalities and biases, and broadens the base of experience upon which the estimates are based. However, it is still an inherently subjective process which does not lend itself well to analysis and refinement.

When applying the analogy approach, the intent is to identify completed projects which have significant similarities to the candidate project, and to draw conclusions about the expected development cost based upon the development costs of the completed projects. This process moves one step farther away from purely subjective opinion, but still does not lend itself to refinement over time.

The parametric approach is the approach most commonly used in automated software cost estimation tools. This approach normally involves a set of basic estimating equations, usually non-linear, where the values of the coefficients and exponents are determined from the application of statistical analysis to a database of productivity data derived from completed software development projects. In addition, a set of factors which affect productivity are identified, and a set of ratings for each factor is developed such that the appropriate characteristics of the candidate project can be quantified. The desired estimates are then computed by some formula which utilizes this set of numerical data. This is a repeatable process inasmuch as providing the same inputs will produce the same outputs, and it lends itself well to refinement over time as the knowledge of the input parameters matures during the life cycle of the project.

THE COCOMO MODEL

In 1981, Dr. Barry Boehm introduced the COnstructive COst MOdel (COCOMO)[1]. The COCOMO was developed using data from a database of 63 projects which were completed between 1968 and 1979. It is a non-linear parametric model whose exponent is greater than unity, indicating that the larger a program is, the more expensive each line of code becomes.

The principal input to the model is the anticipated size of the program to be developed, expressed in thousands of delivered source instructions (KDSI). This size estimate may be divided into two parts - the new code to be developed and the existing code to be adapted for use in the new program. To provide for uncertainty in the new code size estimate, the model requires estimates of the smallest expected size, the largest expected size, and, within that range, the most likely size. A beta distribution is then used to compute the estimated size. For the adapted code, the amount of rework to be required is expressed in terms of percentages of redesign, recoding and integration. Together, the new code and adapted code combine into thousands of effective delivered source instructions (KEDSI), the parameter upon which the effort calculations are actually based.

The original COCOMO contained fifteen parameters called Cost Drivers. These parameters describe properties of the program to be developed, the development team, and the environment within which the development will be done which significantly effect the productivity of the development team. The cost drivers are intended to be mutually orthogonal, meaning that each one stands independent of the others. It is further intended that any factor which will affect program development cost can be associated with one of the cost drivers.

In addition to producing estimates of the total effort and schedule required to complete a project, the COCOMO computes the distribution of the effort and schedule among the project development life cycle phases. The life cycle phases included in the COCOMO are Product Design, Programming, and Integration & Testing. The apportionment of effort and schedule among these phases is based on the contents of the Phase Distribution Tables, a set of tables which contain the percentages of the total effort and schedule to be allocated to each life cycle phase.

Included Models

The original COCOMO model actually consisted of three estimating algorithms. These three algorithms represented three levels of fidelity and were called the Basic, Intermediate, and Detailed models. In the Basic COCOMO, the estimation is performed at the total project level, and no cost drivers are included. The Intermediate COCOMO permits the total project to be decomposed into a set of components, and the cost drivers are included in the equation. The Detailed COCOMO provides for the decompositon of the program down to the subroutine/procedure level, and also includes the cost drivers.

In 1987, a fourth model was defined. This model, called the Ada COCOMO, was developed to accommodate changes in programmer productivity which can result from careful application of the software engineering practices which are supported by the Ada language. However, its applicability is not restricted to Ada program development. It can be the model of choice in any situation where modern software engineering practices are being followed, regardless of the language being used in the implementation.

In addition to these basic estimating models, an extension is included called the Incremental Development Model. This model permits the development of separate estimates for a set of independent intermediate deliveries, and combines these into a total estimated for the complete project.

Development Modes

The COCOMO model identifies three distinct software development modes. These are called the Organic mode, Semi-Detached mode, and Embedded mode.

In the organic mode, relatively small software teams develop software in a highly familiar, in-house environment. The Semi-Detached mode represents an intermediate stage between the organic and embedded modes. This can mean either an intermediate level of the project characteristics or a mixture of the organic and embedded mode characteristics. The major distinguishing factor of an embedded-mode software project is a need to operate within tight constraints. The product must operate within (is embedded within) a stringently coupled complex of hardware, software, regulations, and operational procedures. It is a characteristic of embedded mode projects that changes to one part severely affect other parts. This results in the development typically being more expensive (less productive) than the more independent organic mode projects.

The three modes are represented in the model by different sets of coefficients and exponents in the estimating equations. The different coefficients reflect the differences in productivity associated with the different modes, and the different exponents reflect the different effect that changes in program size have on programmer productivity.

The Estimating Equations

The COCOMO effort estimating equation is

$$MM = \alpha \times KDSI^{\beta} \times \Pi F_i$$

where MM = Man Months
α = Productivity Coefficient
KDSI = Thousands of Delivered Source Instructions
β = Exponent relating productivity to program size
ΠF_i = Product of the Cost Drivers

The schedule estimating equation is

$$TDEV = \delta \times MM^{\varepsilon}$$

where TDEV = Development schedule in months
δ = Schedule Coefficient
MM = Man Months
ε = Exponent relating productivity to program size

Note that the Basic COCOMO effort equation does not include the cost driver term, and that the coefficients and exponents are different for each development mode.

The COSTMODL Program

The COSTMODL is an automated software development cost estimation program. It runs on IBM PCs and compatible computers. It implements all of the COCOMO models except for the detailed model. In addition, it includes a simplified linear model which was developed by NASA at the Johnson Space Center using productivity data from completed NASA projects. This model is called the "Keep It Simple, Stupid" (KISS) model.

Major Features

Major features of the COSTMODL program are its ease of use, its incorporation of multiple estimating models, its ability to support multiple projects and multiple model configurations, and the capability it provides for a user to completely customize the estimating equations to a particular software development environment.

Ease of Use

The most immediately apparent major feature of COSTMODL is its ease of use. It is delivered with an automated installation utility (CINSTALL) which includes a default set of destination subdirectories for the installation. The user is given the opportunity to modify the directory specifications and then CINSTALL automatically configures the destination hard disk, uncompresses the delivery files into the destination directories, checks for any errors during installation, and, if there were none, transfers to the directory into which COSTMODL was installed and executes COSTMODL.

COSTMODL was designed to be immediately useful without requiring that the user read any printed documentation or attend any user training. The program is completely menu-driven, can be controlled from either the keyboard or a mouse, and includes an extensive context-sensitive help system. Immediately upon entry into the program for the first time, the new user is given instructions for entering the help system, and then is led through a brief tutorial on the COCOMO model, usage of the menu system, the function keys, and the help system itself. Complete user's documentation is included with the program, but the user has the option of going directly into the program and using the user's guide only for reference.

Data Management

Once an estimate has been done for a project, that project can be saved in COSTMODL's project library for future use. Many projects can be saved, and many project data files can be saved for each project. The first required input when a new project is being defined is the project name. This name will appear in the project menu on subsequent runs until it is deleted from the library. At any point in the development of a project estimate, the current status of the project can be saved to a file. Whenever that file is recalled on a later run, control will return to the point in the program at which the file was saved, and any desired changes to the project description can be entered at that time.

When a project data file is saved, it is associated with the model which was being used at the time it was saved. However, the file can be imported into a different model. For example, suppose a new project is being considered, and very little is known about its detailed requirements. A rough estimate can be obtained by using the Basic COCOMO or the KISS model, and the resulting estimate saved to a file. Then, as additional information becomes available about the program to be developed, that file can be imported into the Intermediate COCOMO model for refinement of the estimates. The data in the project file will automatically be promoted to the new model, and, once the new information is entered, the new project file will be associated with the model currently in use. This provides the capability to progress to higher fidelity models as one's knowledge of the program to be developed matures. In addition, the retention of a series of project data files throughout the development life cycle of a project makes available a valuable resource for evaluating the validity of the inputs at various points in the development process. These data files can also be useful for identifying those input parameters which are most likely to have adverse effects on the accuracy of the resulting estimates.

In addition to the project data files, additional data files are included which contain the detailed definitions of the models. The coefficients and exponents for the estimating equations, the definitions of the cost drivers and their associated multiplicative values, and the phase distribution tables are all located in data files.

Customization

The most significant feature of COSTMODL which sets it apart from other automated cost estimation tools is its ability to be completely customized by the user. All of the parameters which define the models are available to the user for modification. The most likely parameters to be modified are the coefficients and exponents on the estimating equations. Let us look at why one would want to do this.

The original COCOMO was defined using a database of 63 projects. This set of projects included a mix of business data systems, military real-time control systems, aerospace programs, operating systems, etc, all of which were completed in the 1968-1979 time frame. It is reasonable to expect that, if you are developing transaction processing software for the banking industry in a highly interactive software development environment, the accuracy of the estimates produced by a model would probably be more reliable if the model had been tuned to that specific environment. If a user can obtain good productivity data either from within his own organization or from one which has similar attributes, this data can be used to recalibrate the estimating equations to that environment. This recalibration normally involves only the coefficients on the estimating equations, and can have a significant effect on the accuracy of the estimates produced. COSTMODL provides the capability for the user to perform the recalibration externally and to enter the resulting coefficients (and possibly exponents) directly into the proper input forms, or he may provide the productivity database directly to COSTMODL and it will perform the desired calibration.

The set of cost drivers can also be modified or totally replaced. The number of cost drivers, their names, attributes, descriptions, and multiplicative values are also under the control of the user. It is not unlikely that a particular development environment contains a factor which significantly affects productivity but which is not included in the original set. For example, suppose a project is being developed in a very large organization and different parts of the program are being developed at widely separated locations. This can affect the total productivity, but is not included in the default cost driver set. It is simple in COSTMODL to add a cost driver for that factor.

It is possible that different organizations may include different activities in the various development life cycle phase definitions, resulting in percentage distributions of effort and schedule which do not agree with the original COCOMO definitions. In this case, the user can modify the contents of the phase distribution tables to reflect those differences, resulting in estimated phase distributions which more accurately reflect those expected within his organization. In addition, the next release of COSTMODL will provide the capability for a user to change the number and names of the life cycle phases in addition to the percentage distributions.

Since the behavior of the program is completely dependent upon the contents of the various data files, it is important that the integrity of these files be carefully protected. To provide for the situations where multiple users may use COSTMODL on the same machine, or when COSTMODL is installed on a network server for use by many people, that portion of the program which permits modification of the configuration files is password protected. This makes it possible for the system administrator to control the contents of these files, and all of the users to make use of the configurations that have been standardized for use within that organization.

Incremental Development Model

The COSTMODL implementation of the incremental development model provides the user with complete flexibility in the definition and modification of the increments, ordering them, and relating one to another. Each increment is created as a total project and saved to a file. Once all of the increments have been defined, the user identifies them in sequence to the incremental development model. As each increment is added to the project, the user specifies the time relationship between it and the previous increment. For example, the Critical Design

Review (CDR) for the fourth increment might be scheduled to be held two weeks after the beginning of coding on the third increment.

In addition to the time relationships between the increments, the user specifies the expected amount of rework to be done on the previous increments to accommodate the next increment. This adaptation effort is frequently called the "breakage factor".

The incremental development model computes the total effort and schedule required to complete the total project, and, for comparison, computes the effort which would have been required had the total project been delivered as a single delivery. In addition, the cumulative staffing level required as a function of time is also computed. The overall schedule can be displayed graphically showing each increment and each life cycle milestone within each increment. The cumulative staffing level can be similarly displayed in the form of a histogram. To facilitate relating the staffing changes to the beginning and ending of the various increments, the two plots can be superimposed.

Each of the plots can be output in the form of a printed report. The total time duration and the number of increments is examined to determine the page orientation and number of pages to be used for the printed displays. It is therefore possible to print complete schedule and staffing plots for any size project.

INDUSTRY ACCEPTANCE

COSTMODL has enjoyed broad acceptance within the software cost estimation community. It is currently being used at more than 300 installations throughout the aerospace community, government agencies, DOD, and academia in the U.S., Europe and Canada. It is a principal estimating tool in use at NATO headquarters in Brussels. It is also being used as a teaching tool at several universities, and was recently selected as the best overall implementation of the COCOMO model in a competitive evaluation conducted by MIT's Sloan School of Management.

COSTMODL is also being distributed to the students and staff at the Defense Management Systems College, and is being distributed by the Air Force Cost Center through their software and data distribution channels.

TYPICAL USES

Typical uses of COSTMODL include feasibility assessment, bid preparation, bid evaluation, budget/manpower planning, and project scheduling. In a time of severe budget constraints, a defensible early estimate of the cost to develop a candidate software product can help avoid spending resources on a project that may not be cost effective. It can also be quite valuable in the control of cost overruns. What is seen as a cost overrun is frequently the result of an underestimate of the effort which can be expected to be required.

The use of a formal cost estimation tool in bid preparation is obvious. It can also be quite valuable to the team evaluating a proposal. It is at this point that possible underestimates can be identified, resulting in the possible avoidance of either perceived cost overruns or even failed projects.

A less obvious benefit resulting from the use of a parametric model can be the thorough documentation of all of the assumptions that went into the development of the estimate. Suppose funds are being sought for a new program development, and senior management is challenging the estimated cost to complete the project. The program size estimate and each of the cost driver inputs can be discussed to see if there is disagreement with any of the inputs, and, if changes are agreed upon, a revised estimate can be produced. Further credibility can be attached to the estimates by referencing the productivity data in the database which was used to calibrate the model, particularly if in-house data was used.

APPLICABILITY

A tool such as COSTMODL can be an important asset for any organization which has a significant financial interest in software development cost management. This includes organizations which develop software products for the commercial marketplace, organizations which develop custom software, the organizations which contract for the development of custom software, and organizations which develop software in-house for their own use.

FUTURE PLANS

COSTMODL will continue to incorporate the latest advances in software cost estimation technology as they become available. The next version will include the capability for the user to redefine the software development life cycle by adding or deleting life cycle phases, or modifying the definitions of existing phases. The classifications of personnel to be included in the development cost computations can be defined, along with their dollar cost per man month and the percentages of their effort to be applied to the current project. An expanded adapted code model will be included which will better accommodate the situation where a new product will consist mainly of adapted code, with a relatively small amount of new code to be written. This situation is quite common in a mature organization where a large pool of software exists which is specific to that organization's function, much of which is adaptable for use in new software products.

Software sizing models are being investigated for possible incorporation. A sizing model allows a user to describe the product to be built in functional terms instead of thousands of lines of code. Function Points are proving to be useful in business applications, but have not been as successful in the scientific software domain. Examples of function points are display screens, input forms, number of input fields, number and size of external databases, etc. COSTMODL will incorporate a Function Point sizing algorithm for use on appropriate applications. The evaluation of candidate sizing algorithms for scientific applications will continue.

The COSTMODL developers are currently involved in studies intended to expand the estimating models to include Fourth Generation Language development, Knowledge-Based Systems, and Totally Reusable Software Libraries. Each of these areas are increasing in importance, and are not well modeled by current algorithms.

Another effort currently underway is the integration of COSTMODL into a total project management system. This integrated capability would combine the effort and schedule estimating features of COSTMODL with the real-time collection of actual expenditures of resources during the development life cycle of a project to continuously refine cost-to-completion estimates.

AVAILABILITY

COSTMODL has been submitted to NASA's Computer Software Management and Information Center (COSMIC) for distribution into the private sector. COSMIC can be contacted at:

The University of Georgia
Computer Services Annex
Athens, GA 30602
(404) 542-3265

In addition, copies can be requested through the Software Technology Branch Help Desk at (713) 280-2233

1. Boehm, B. W., *Software Engineering Economics*, Englewood Cliffs, NJ, Prentice-Hall, 1981.

N92-22710

INCREASING PRODUCTIVITY THROUGH TOTAL REUSE MANAGEMENT (TRM)

M. P. Schuler
NASA Langley Research Center
Hampton, VA 23665

ABSTRACT

Total Reuse Management (TRM) is a new concept currently being promoted by the NASA Langley Software Engineering and Ada Lab (SEAL). It uses concepts similar to those promoted in Total Quality Management (TQM). Both technical and management personnel are continually encouraged to think in terms of reuse. Reuse is not something that is aimed for after a product is completed, but rather it is built into the product from inception through development. Lowering software development costs, reducing risk, and increasing code reliability are the more prominent goals of TRM. This paper describes procedures and methods used to adopt and apply TRM. Reuse is frequently thought of as only being applicable to code. However, reuse can apply to all products and all phases of the software life cycle. These products include management and quality assurance plans, designs, and testing procedures. Specific examples of successfully reused products will be given and future goals will be discussed.

WHAT IS TOTAL REUSE MANAGEMENT

Total Reuse Management (TRM) is achieved when an organization as a whole and its individual members make a commitment to continuously: search for reuse opportunities associated with each work product and activity; distribute notices to appropriate members of the organization of identified reuse opportunities; and actually reuse products, ideas, methods, and procedures. It is similar to Total Quality Management (TQM) in that its success depends on individuals adopting a mindset which influences all their activities, that is, reuse becomes a way of thinking and acting. With the expectation of increased productivity, management provides the necessary time and personnel for the record keeping and documentation required to support TRM reuse activities.

GOALS OF TOTAL REUSE MANAGEMENT

The major goal of TRM is to reduce software development costs by reusing previously developed knowledge and products instead of redeveloping them. Goals directly related to code development include reducing risk and enhancing reliability. The growth of reuse within a project decreases the number of elements to be developed which decreases the stress on the project schedule and thus lowers the risk of overruns. Since each time a component is reused it is retested in its new application, the additional error discovery and correction process naturally yields an enhanced product reliability. The greater the number of reused components the lower the overall risk of mission failure. Further, the successful demonstration in a previous project of a proposed method of implementation, increases confidence that the method is a worthy candidate for reuse. Another goal is process improvement since reuse can include procedures and methods as well as products. By documenting the improvements in procedures followed on a given project those improved techniques become available for reuse on future projects, thus improving the organization's procedures and methods over time.

THE TRM ADOPTION PROCESS

The first step is to instill a reuse mindset into the members of the organization. This was done from the top down at the Software Engineering and Ada Lab (SEAL). The SEAL adopted the idea of reuse as one of the lab's primary objectives. To achieve this objective, the regular weekly software management staff meetings were supplemented with discussions to promote reuse, determine individual project components which had reuse potential, and focus on reuse successes. Each project manager would report on products, concepts, and ideas which they considered candidates for reuse. In addition, project managers were required to report on products which had been successfully reused. Listing examples of these products gave other project managers ideas on where reuse might be possible on their project. Once the project managers began to adopt a reuse mindset the weekly reports were changed to monthly ones. Similarly, project managers were responsible for the introduction of TRM on their individual projects at project level staff meetings. In addition, a catalogue of various types of products reused was kept and is now being used to aid in training new employees so that they can quickly become contributors to the TRM program.

TYPES OF REUSE

The following sections give a variety of different areas and products in the software life cycle where reuse can be effectively established to lower development costs, reduce risk, and increase functional reliability. Reuse can occur across organizations, across projects, and within projects. These examples are given to illustrate the multiplicity of reuse situations which yield substantial benefits and to stimulate ideas concerning the applicability of reuse concepts at all organizational levels.

Reuse Through Technology Transfers Across Organizations

The Formal Inspection process is comprised of a set of procedures designed to efficiently detect and correct errors. Jet Propulsion Laboratory (JPL) had already developed a Formal Inspection Program and had years of practical experience with the process when the SEAL decided to adopt it as one of its measures to improve software product quality. JPL agreed to perform a technology transfer of information and products to the SEAL to support the initiation of the program. This saved the SEAL the hundreds of man-hours it would have cost to start up the program on their own. The SEAL was able to reuse JPL's managerial framework, training course materials, checklists, forms for conducting inspections, forms for recording inspection statistics, and a copy of their customized software to run the data base for statistical collection of inspection results. JPL also provided averaged cost estimates by software development phase for man-hour expenditures on inspections. By reusing their cost data the SEAL was able to calculate the estimated cost of inserting the inspection procedures into new projects. In addition, the two organizations agreed to communicate regularly to exchange information, documents, and data that could be reused. This is important since it assured that technological gains made at each site would be exchanged and reused.

JPL also trained SEAL staff members to teach their training course on formal inspections and those members are now actively involved in the process of conducting the same technology transfer to numerous organizations within NASA as well as different companies in industry. Not only are the formal inspection products and procedures being reused but the methods used to insert the process into the organization are also being documented and reused to further facilitate the transition. For example, for in-house projects, a set of procedures for introducing formal inspection has been established and is reused with each new project interested in inspections. For external projects, a notebook containing an insertion plan and all records of the execution of that plan is being compiled. All lessons learned from the insertion process and references to information used in the process, such as cost estimates, will be recorded in the notebook. Records tracking such things as number of inspections and actual cost will also be included. This information will be reused on future external projects.

Object Oriented Design (OOD) was another area of technology which the SEAL group was interested in. As the group recognized the need for incorporating OOD in its practices, the services of an organization expert in consulting and teaching OOD were employed. This substantially reduced the cost and time required to implement OOD technology within the SEAL. Once the insertion process was complete, an agreement was worked out between the expert organization and the SEAL group to exchange lessons learned and copies of completed designs for reuse on future projects. In addition the SEAL has developed an informal network to exchange information on OOD between several organizations and is now in the process of receiving reusable case study feedback from network members.

In both of these examples, Formal Inspections and OOD, the SEAL was successful at completely reusing technologies from outside organizations to improve its own software engineering capabilities. In addition, a core group of participants was established for each new technology to handle the future distribution of related reusable materials among the different organizations participating. This has resulted in an increase in productivity in two ways: by exchanging lessons learned and advancements in procedures followed by the individual organizations the core group has been able to achieve process improvements at a relatively rapid pace and low cost; and, by directly reusing completed products, considerable manpower has been saved. In addition the Seal makes it a point to regularly have informal technical interchange meetings with other organizations on specific areas of mutual interest. Such areas include: Systems Engineering, Object Oriented Design, the use of Ada programming, reuse, and risk

management as well as specific applications domain work. These meetings are conducted to transfer reusable technology between NASA sites and from NASA to Industry. The SEAL encourages and invites this type of interaction with all organizations interested in participating in technological transfers.[1]

Developing Expertise for Reuse within an Organization

Each individual in the SEAL organization was spending a considerable amount of time installing and becoming proficient in the use of new computer devices and software tools. A method was devised to reduce the overhead associated with these learning activities. For each new tool that was purchased, one or two individuals were given the responsibility of becoming experts on it. This involved searching through the technical reference manuals, learning how to set up the tools, and becoming proficient at the most common commands needed for typical use. Once this was completed, the experts reused their knowledge to install the tools on each computer and gave a brief seminar and list of common commands (along with explanations) to the general users to acquaint them with the tool in an expedient manner. This process saved the organization a considerable number of man-hours and was used on numerous devices and software tools such as: compilers, debuggers, logic analyzers, PC-based tools, PC-based networks, and specific bus protocols. If a tool was too complex to learn in a short seminar, the tool vendor was brought on-site to teach SEAL staff members how to give a detailed instruction course on the tools spectrum of functions and applications. Masters of the instructional materials needed to teach the courses were provided with the training. In depth classes on these tools are now given regularly with the only cost being the SEAL staff members salary and the cost of reproducing course materials.

The SEAL is also involved in bringing on-site a number of software engineering courses that are taught on a one time basis. A catalogue of course titles and attendants is kept so that if the need arises knowledge in that subject area can be easily tracked for reuse. Apprenticeships can also be set up to transfer knowledge for reuse within the organization. For example, one of the designers took on an apprentice at the beginning of a subsystem design and by the midway point the apprentice was a constructive contributor to the design. When the project was completed the apprentice took that newly gained expertise and reused it to design a subsystem on another project.

Finally, if training tapes were available from vendors they were purchased and reused throughout the organization. The cost of travel and tuition for technical training courses can put a substantial drain on an organization's budget. By reusing expertise, instructional materials and tapes, an organization can become educated in new software engineering technology at a much lower cost.

Reuse Within And Across Software Projects

The following is a description of how actual software products are being reused both within and across three different SEAL projects. All three projects are written in Ada and use Object Oriented Design Methodologies.

Project I[2]

Communication drivers are the software components developed to drive bus communication and they are among the most time consuming components to write. Designing them to be reusable would yield significant savings on future projects. This requires the use of a design approach which is both modular and hierarchial. A module is developed to serve as an interface between the applications level code and the code responsible for driving bus communication. The interface module is designed to provide a general purpose (application independent) interface to the bus while the associated bus modules are written to perform actual bus communications for a specific standardized bus protocol on a specific manufacturer's card. This promotes reuse in two ways. First, the applications level modules can be reused without modification if the bus should be replaced with one of a different type, since they had been developed with no dependencies on the type of communications hardware. Second, the software written to

1. For more information on technology transfers or training seminars contact Pat Schuler (804) 864-6732.
2. Copies of the code mentioned in this section are available courtesy of the Controls Structures Interaction Project. Contact Pat Schuler (804) 864-6732.

drive the bus communications can be reused on any project containing the same supporting communications hardware since the interface module was written to be applications independent.

Both of these types of reuse have been achieved on SEAL projects. For example, Project I used a MIL-STD 1553 bus in its PC-based Ground Support Equipment to test and record 1553 communications from a flight computer. The Ground Support Equipment, and over 4,000 lines of Ada code for 1553 communications, were reused to produce a similar checkout subsystem for a completely different flight project, Project II. In addition, the design for Project I was reused on a project in the same applications domain but with different hardware and software language requirements. This was possible since reusability had been a principle design requirement which assured that the completed product would not be dependent on the target hardware.

There were other types of communications software reuse within the same project. Project I contained three major subsystems and all three communicated across the 1553 bus. Subsystem I, used the 1553 bus to configure Subsystem II with information necessary to perform normal operations. Subsystem I was an INTEL machine and Subsystem II used a 1750A computer but the data structure used to hold the configuration data was identical on both systems. The data structure was developed on Subsystem II and reused on Subsystem I. Also, the data structures used to implement the remote terminal mode of the 1553 communications were reused without modification and the Ada package specification for both the bus control and remote terminal modes were reused between the subsystems with only slight modifications.

Another software component that was reused was a special purpose block move which was developed for Subsystem II. Its purpose was to increase throughput by transferring vectors rather than words of data from the 1750A to the array processor card and from the 1750A to the 1553 card contained within Subsystem II. This software was reused to improve communications rates within Subsystem III between its 1750A and digital signal processor card and also between its 1750A and 1553 card. Also, Subsystem III had an additional requirement. It not only had to accept configuration data from the 1553 bus but must be configurable via an RS232 bus. The code for reading in the user supplied information and processing it so that it could be used to set the configuration was reused from Subsystem I to fulfill this requirement for Subsystem III. This saved approximately 4 man-months of effort.

The ease with which the 1553, the block move, and the configuration code were reused is in a large part due to the use of Object Oriented Design (OOD) and Ada. OOD methods made it possible to minimize the communication between software modules while maximizing the cohesion between routines performing functions related to specific objects (or devices) in the system. The use of Ada constructs such as packages supported these methods and the portability of the language made reuse across different subsystems with different processors and bus interfaces possible. Further, each software component was required to use a standardized prologue stating specifics about the software it contained: compiler options, author's name, and date, etc. There were standards for how the software was to be written and formatted as well. Both of these contributed to readability which made software components easier to understand and therefore easier to reuse or modify for reuse.

Project I also reused, from another flight project, part of a math library written in assembly language. Although it had to be converted to 1750A assembly code the algorithms were reused. During the conversion process an error in one of the math routines was discovered and fixed on the 1750A code as well as the original source code. By reusing software across projects and subjecting it to different environments and applications, previously uncovered errors are found and fixed. The risk of future failure is decreased since its test domain has increased and that is a major advantage to reusing code. Its reliability increases the more it is reused. In the same way, when the 1553 code developed on Project I was reused for a different application on Project II, errors were discovered and the 1553 code was fixed on both projects and its reliability was increased as well.

System tests procedures can be reused as additional subsystem tests to eliminate errors prior to integration testing. This was accomplished by connecting Subsystem I and II with a simulator for Subsystem III which was not yet completed. These subsystems had already been individually tested, but performing the system integration tests on the hybrid system acted as an additional suite of tests and uncovered errors which would not have been discovered until system integration testing. Further, since the hybrid system had now been verified, when errors did occur during full system integration testing, the errors were contained to either the newly added subsystem or communications

with that system. This substantially reduced the time spent in debugging. Therefore, reusing system integration test procedures as additional unit tests can flush out problems early and help isolate problems as each new subsystem is added.

Project II[3]

Project II reused the standardized prologue developed on Project I which insured consistency in software component documentation across projects. The Ground Support Equipment and the Mil-STD 1553 bus code were also reused from Project I. The 1553 bus communications drivers were reused without modification because they had been developed with a general purpose interface module as the delimiter between the applications specific software and the hardware specific software. This ability to reuse not only the code but the actual computer equipment provided Project II with a substantial savings.

Project II achieved even better results with code reuse between its subsystems. The project set up a directory where completed code was placed so that it was available for all project members to reuse. Among the products stored in this directory are software components to perform: windowing capabilities; text and menu displays; and a general purpose parser for converting transmission packets. In addition it also contains software for; UART / serial communications via the RS232 protocol, queues, ring buffers, semaphores, data structures, and time packages which provide information such as the julian date and system time. Each of these components are being reused within Project II across several subsystems and this has substantially reduced the total cost of the system being developed. Project II has produced a substantial amount of code for INTEL hardware which is available for reuse. A complete Ada interface for all registers and modes has been written for the following; INTEL 8259 Programmable Interrupt Controller, INTEL 8254 Programmable Interval Timer, INTEL 8237 Programmable DMA Controller, INTEL 8255 Programmable Peripheral Interface, and INTEL 8274 Multi-Protocol Serial Controller. In addition, a complete Ada interface for all internal registers of the INTEL 80186 microprocessor has been written and is available for reuse.

Project III

This is the third project started since the SEAL began the TRM program. It will be reusing the 1553 bus components and the coding prologues that were developed on Project I as well as the coding standards from Project II. In addition, a substantial amount of code will be reused from Project II such as software to perform; windowing capabilities, text and menu displays, keyboard input manipulation and filtration, and RS232 protocol. It will also reuse packages which provide functions to; manipulate linked lists, implement the semaphore construct, declare varying length strings, log events with time stamps, and replay prerecorded user input for batch testing. It will also reuse many of the software components developed for INTEL chips mentioned above.

It is clear from the three project examples that a substantial amount of reusable code can be accumulated in a relatively short amount of time. However, it is important to recognize that the successful reuse of code on these projects was largely due to developers regularly reporting on the completion of potentially reusable software components. Even more important was the staff's commitment to build each component to be reusable. To facilitate this, modern design techniques such as those found in Object Oriented Design were used along with the Ada programming language which provides constructs to support those design methods.

Software Development Documentation Procedures That Encourage Reuse

Software documentation is a major expense on any software system. The key to reducing that cost is to use methods and procedures that result in development documents which can then be reused for formal reviews and for the required deliverable documentation. This can be done at each phase of the software lifecycle. During requirements analysis phase event sequences and responses, data elements, timing order and constraints as well as states and processes can all be documented in tabular, graphical, or textual form as part of the stepwise analysis

---

3. Copies of code mentioned in this section are available courtesy of the Lidar In-Space Technology Experiment. Contact Chuck Carpenter (804) 864-8046.

process. These documents are the output products of the requirements phase. If careful attention is paid to thoroughly documenting each step of the process those documents can be reused, without alterations, in the requirements document and for the requirements review. Once the requirements phase is complete, object diagrams with written object descriptions and operation descriptions as well as a requirements traceability matrix and decomposition tree showing the hierarchical connections in the design can be used to thoroughly document the activities required to complete the preliminary design phase. These output documents, again without modification, are reused as the deliverables for the preliminary design and the preliminary design review. If all of these documents are routinely updated as part of the activities for the detailed design and coding phases, they may also be reused not only for the detailed design review and test readiness review but could be reused for the 'As Built Configuration Document' which describes the final product. In addition, automated testing procedures along with confirmed test results can be reused in the users/operators manual as diagnostics. The 'As Built Configuration Document' and the test procedures can also be reused during maintenance phase. They are invaluable to the maintainers as a record of the system's purpose and how the functional specifications were achieved. As modifications are completed these documents are updated so that they can be reused for the life of the maintenance phase.

In addition to the above-mentioned reuse benefits within a project, these documentation techniques encourage reuse across projects. Since the products are well documented and the documents are kept current, other projects will find it easier to understand the product's requirements and functions. This ease of understanding makes the chore of locating reusable components less time-consuming and therefore more attractive. In addition, since not only the code but the requirements and design are available for each component, savings from reuse can also be achieved during requirements and design phases on new projects.

For this type of document reuse to be achieved a defined process with specific steps and procedures must be in place before a project starts. These procedures need to specify the form and content of all output products for each step and their function as input documents into the next phase. Instead of the developers producing large quantities of documentation that they incorrectly perceive as the customer's needs, a true picture of the current project posture can be obtained from reusing the actual development documents as the review documents.

Reusing Documents From Other Organizations And Projects

When the reuse program started, the SEAL had no documentation of its own to reuse. Therefore, an effort was made to obtain existing documents from numerous organizations within government and industry. A documents library was started and currently holds project documents covering almost every phase of the software lifecycle: Software Requirements, Interface Requirements, Test Procedures, Management Plans, Configuration Control Plans and Verification and Validation Plans. Portions of these are reused each time a new document for a software project is written and this has saved a considerable amount of man-hours. Procedures for constructing testing trees or a traceability metrics, for example, need only be written once and can then be reused in the documentation for other software projects. The library also holds documents on how to perform specific software engineering tasks such as quality assurance, design, benchmarking compilers and a variety of more general documents on software engineering standards and procedures. All of these are reused on SEAL projects to determine various software development procedures. In addition, several contracting organizations responsible for developing software for the SEAL have also reused many of these in developing deliverable software documents and establishing their own software procedures.

Standardizing On An Environment Promotes Reuse

The SEAL has a standard environment for software development which includes; computers, compilers, debuggers, logic analyzers, etc. The SEAL has also standardized on word processors, drawing tools, spreadsheet programs and data base managers. To make reuse easier, each staff member is connected to a local area network. Network licensees were purchased for the tools in the standardized environment. That proved to be more economical than purchasing individual copies. Status reports and presentations can be constructed by reusing information electronically obtained from the various nodes on the network. The tools are chosen based on their compatibility to support such activities. Forms are constructed using these standardized tools, for documenting common activities

such as purchase requests and travel requests and inventorying books, software and hardware, etc. Those forms are made available to all users of the network to reuse.

FUTURE DIRECTIONS

Once documentation is completed, all the SEAL software components will be submitted to COSMIC (the NASA reuse library) for general access. Project data such as man-hours, lines of code and pages of documentation being collected on current projects will be reused. The data will be entered into existing off-the-shelf estimating tools such as PRICE and REVIC. Entering actual data from SEAL projects into tool data bases will customize the output to the SEAL environment and potentially give more accurate project cost estimates. Even more extensive metrics will be recorded on future projects and will be reused to perfect further project estimates. In addition, descriptions of completed code written for specialized hardware used on SEAL projects will be given back to the vendors of those hardware products. They will be encouraged to relay this information to their customers via newsletters or computerized bulletin boards so that organizations purchasing those hardware products may have the opportunity to reuse completed SEAL code. The SEAL also has plans to develop its own local library of reusable components to hold software for such things as ring buffers, stacks, math library routines, etc. Software components for specific chips and busses will also be catalogued and entered in the library. The SEAL is currently promoting standardization on specific busses and chip sets. Reuse can be maximized by using standardized hardware and reusing existing custom-built software and test environments to substantially cut future project costs. A software project handbook will also be written to provide information on established methods and procedures for reuse on future projects. In addition, work is currently being done to develop a set of generic management, quality assurance, testing, etc., plans that will act as templets to be reused by all future projects and project specific information will be inserted to customize the plans. To promote reuse on a center-wide scale, a Software Engineering Users Group is being started which will meet regularly to exchange information on software methods, procedures, tools, and completed modules.

CONCLUSION

If the current activity will ever be performed again, there is potential for reuse. Perform the activity accordingly so that it or its products can be easily reused and document its existence so others can locate, understand, and reuse it.

N92-22711

THE USE OF HYPERMEDIA TO INCREASE THE PRODUCTIVITY OF SOFTWARE DEVELOPMENT TEAMS

L. Stephen Coles
Group Chief Technologist
Information Systems Integration
Institutional Data Systems
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099

ABSTRACT

Rapid progress in low-cost commercial PC-class multimedia workstation technology will potentially have a dramatic impact on the productivity of distributed work groups of 50-100 software developers. Hypermedia/multimedia involves the seamless integration in a Graphical User Interface (GUI) of a wide variety of data structures, including high-resolution graphics, maps, images, voice, and full-motion video. Hypermedia will normally require the manipulation of large dynamic files for which relational data base technology and SQL servers are essential. This paper will cover basic machine architecture, special-purpose video boards, video equipment, optical memory, software needed for animation, network technology, and the anticipated increase in productivity that will result from the introduction of hypermedia technology. It is suggested that the cost of the hardware and software to support an individual multimedia workstation will be on the order of $10,000.

INTRODUCTION

Extrapolating the ten-year history of PC-class personal computing, we can forecast that the next decade will witness progress at an accelerating pace. By the year 2001 we can expect to buy a "4G" workstation [where the "G" stands for "giga" as in 1 GIPS, 1 GB RAM, 1 Gpixel $(30K)^2$, 1 Gbit/sec BW] for about the same price as we pay for an ordinary PC today. What will we be doing with such a supercomputer on our desktop? Obviously, we will be solving problems for which this sort of power will be taken for granted in the same way that no one would ever return to a conventional typewriter and carbon paper after using a PC word processor plus copy machine to get routine work done in the same way that practitioners of the old typewriter technology would never go back to using ink quills and parchment for written communications commonplace in the time of their predecessors.

The key to improved productivity in the next Century will depend on hypermedia. By definition, hypermedia is the seamless integration in a GUI (Graphical User Interface) or Windowing Environment of a variety of data structures, including: numbers, text, mouse-clickable icons, 3-D bit-mapped color graphics (charts, graphs, etc.), animation, maps, high-resolution images (e.g., Landsat Images), voice, music, full-motion video, and what has now come to be called "virtual reality." Each level of detail in a complex representation of an information structure will be "mouse clickable" up to the maximum level of resolution possible in whatever manner the user wishes to access it. Sometimes a cursory level or abstraction is adequate for understanding and extra detail would only clutter the landscape, while at other times the maximum level of detail is indispensable for understanding. Since the author of a complex information structure cannot anticipate the needs of each user, the information must be presented at the highest level of abstraction in such a manner that the user has the option to interactively "mouse click" his way into the supporting data as he sees fit. Hypermedia may not be ideal for all forms of communication. For example, a literary novel is essentially a flat data structure (with the exception of occasional chapter headings). However, the design, construction, and operation of an advanced spacecraft is a multi-hierarchical activity to which a user cannot usefully contribute without the selective access to an extraordinary array of detail. Compression of detail with selective accessibility when needed by using hypermedia would be an ideal way for an engineer to come up to speed quickly on a large body of information.

Thus, hypermedia will involve the transparent manipulation of large distributed data bases that must be updated regularly, if not in real time (e.g., telemetry). Such updates may also involve uncertain or even contradictory data derived from disparate sensory modalities.

To achieve the feeling of presence, a team of 50-100 software developers, working in different parts of the world (in different time zones), not only need to be able to transmit reusable object-oriented programs to one another over a common network as they do now; they must also have the ability to transmit e-mail messages containing mouse-clickable icons of short bursts of full-motion video that show how a program actually behaves during operation together with mouse clickable icons of digitized voice annotation commenting on these programs. Remote debugging of jointly developed code through a combination of e-mail, remote login, and telephone communication does not always provide the sensation of presence needed to expedite the process of debugging. Using today's technology the process of debugging when the author of a program is physically present vs. interacting on the telephone can be an order of magnitude faster. In the future using hypermedia and desktop teleconferencing over ISDN telephone lines, the requirement for physical presence will be completely obviated. Furthermore, the best members of a software team can be rapidly assembled to carry out a short development task or form a "tiger team" to debug a problem during a crisis, regardless of geography. Indeed, at a more distant future time telepresence will eliminate the need for commuting to one's job, and one could conceivably live in any location one desired.

Other applications for hypermedia technology include the rapid creation of S-VHS video tapes (not broadcast quality) for executive presentations, ordinary briefings, and training. Indeed, one of the major untapped areas for this technology is in education. The only thing educators are waiting for is for the cost of the technology to come down sufficiently to justify it as a viable alternative, which will certainly happen in the next few years.

YOUR OWN MULTIMEDIA/HYPERMEDIA WORKSTATION

In this section we will provide our "recommended list" for the acquisition of parts and assembly of a standard multimedia workstation. This includes the following topics: basic machine architecture, special-purpose boards to handle full-motion video, commercial video equipment, optical memory, animation software, networking, information retrieval from large data bases, integrating sound, and integration and interoperability.

Basic Machine Architecture

The choices include various PC clones (DOS), Apple Macintosh (System 7), Amiga, IBM PS/2 (OS/2), and Compaq 386/33Mhz or 486. The estimated cost for such a PC clone is in the neighborhood of $4K. More expensive workstations are also suitable if one already has such a machine that can be dedicated to this function. These include engineering workstations such as Sun Sparc, DEC Station (MIPS), HP 7000, IBM 6000, Motorolla 88000, NeXT, or even a special-purpose graphics workstation like a Stardent or a Silicon Graphics.

Recommended examples of standard PC Options are as follows: 8MB of RAM Memory plus 100 MB or more of hard drive; an NEC MultiSync 4D or better high-resolution color monitor; a super-VGA Graphics board such as Video Seven VRAM-II; a Hayes-compatible modem; a Western Digital ethernet board, a two-button mouse (cordless); dual floppy drives (3 1/2" and 5 1/4"); enhanced keyboard (101 keys); one parallel and two serial ports; DOS 5.0 and Microsoft Windows 3.0. A math co-processor is optional. Access to a Laser Printer is assumed.

Special-Purpose Boards

A frame grabber or image capture board is essential (Super Via from Jovian Logic for PCs or VideoPix for Sun SparcStations). A scan-converter board, such as the VGA Producer from Magni Systems, is needed to convert RGB graphics to NTSC television format. This provides a capability for graphic overlays, like icons positioned on a map. A full-motion video in a window board, such as Super Video Windows from New Media Graphics, is needed. This allows any video source (Live Camcorder, Over-the-air or Cable, VCR, or Laser Disk)

as input in a fully-scalable, movable window. An audio card like the SoundBlasterPro for output or IntroVoice VI for input is also useful. The estimated cost of these boards is about $2K. Optional cards include a FAX board (FAXPRO) or a VCR Controller (Diaquest), but they do tend to bring up the price.

Other useful display options include an LCD flat-plate color projection screen for an audience of 5 to 30 persons, a color hardcopy printer, and a scanner (for example, the HP Scan Jet flatbed scanner with accompanying OCR software and/or the A4 hand-held color scanner) are useful. In 1993 Intel plans to introduce its DVI chip set for video compression (temporal and spacial) and this will represent a significant compression in bandwidth needed to send video images over leased phone lines.

Consumer Electronics Video Equipment

A high-resolution color monitor, a Hi8 camcorder (Canon) and tripod, two s-VHS VCRs (capable of independent control of luminance and chroma levels), a Laser Disk Player (Pioneer 4200), and two active stereo speakers (Sony bookshelf @ 30 watts each) are the minimum needed to get started. The estimated price is $2K. Optional VHS editing decks, such as from Sony or Panasonic, are needed to handle complex animation sequences and will simplify the video editing task.

Flat screen, 50" on a side, wall-mounted HDTV displays will come by the end of the century. Preliminary HDTV laser-disk experiments at consumer electronics shows proves that their quality will be as good as film.

Optical Memory

A CD-ROM player attachment for your PC can now be bought for under $500. Many publishers are now considering this as the medium of choice for periodicals, images, reference works (encyclopedias, dictionaries, atlases, classical literature, etc.), and now even software (operating systems and third-party applications are now distributed on CD-ROM).

Sectorized erasable magneto-optical media storing 640 MB of data per cartridge ($120 each) are now routinely available and are rapidly making WORM (Write Once Read Many) media obsolete. The optical density of these cartridges is soon expected to be doubled. Furthermore, the creation of jukebox technology through a SCSI interface at reasonable prices (a 10-platter model from IDE gives a total of 6.4 GB for about $10K while a 144-platter model from HP gives a total of 120 GB for about $80K) gives a cost per bit of storage at reasonable access times that can't be beat by hard drive technology.

Animation Software

The choice of operating system (DOS 5.0, UNIX, OS-2, or System-7) is dictated largely by the vendor of the hardware platform. However, as soon as a standard version of UNIX becomes more popular at the PC (Solaris 2.0, LAN Workplace) and MAC level in the next few years, this will be increasingly less so. A windows environment (Windows 3.0 or Microsoft New Technology [NT] Operating Environment, SunView, X-Windows [OpenLook 2.0, Motif, X-Vision, DeskViewX]) is needed to support a mouse for "point-and-click" and "drag-and-drop" operations. In addition, standard development tools (Omnis-5, SuperBase-IV, TookBook, VisualBasic) and standard applications for word processing/desktop publishing/office automation (Frame, Microsoft Word, WordPerfect), spread sheet (Lotus 123, Excel), project management (Primavera, Project), Expert System Shell (Kappa, ART), and relational database systems (FoxPro, Oracle, Sybase) will be needed. The cost of these packages is not included, since it is assumed that one would be buying these anyway.

However, a number of commercial software packages have been created to facilitate the overlay of animated 3-D graphics on full-motion video integrated with sound (voice narration/music), including GRASP 4.0 from Paul Mace Software, SantaFe Media Manager from HSC Software, MacroMind Director, 3-D Studio from AutoDesk, and the Video Toaster from NewTek, Inc. (special effect library). These packages have stretched the English language with their own new vocabularies (e.g., masking, keying, embossing, posterization,

mosaicing, pixelation, voxelation (volumetric pixel), chiseling, texture operators, transparency/translucency/opacity/specularity/gloss/matte/diffuse, shading (Phong/Gouraud), "lofting" (extrusion), morphological operators (erosion/dilatation), "tweening", "morphing" (3-D), sparkle wipe, fade, dissolve, shrink-and-tumble, etc.). The estimated cost for this software is $1.5K.

To my knowledge, no software package is yet available that addresses the need of amateur video producers to compute automatically the number of video frames required to cover a spoken sound-byte (previously recorded with a known duration in milliseconds of digitized speech) so that it can be spliced in (drag and drop) at the click of a mouse. Conversely, visually sliding a music clip over a certain number of frames with the proper fade in/out characteristics mixed in with voice narration would be a substantial productivity boon for the home video editor.

Networking

Client/Server architecture is increasingly popular in Local Area Network design, since it distributes computing cycles appropriately over a network. Network communication packages (such as Novell 386, Banyan Vines, LAN Manager, PC Plus, CrossTalk) and e-mail (DaVinci) are needed to manage network administration. However, for nationwide workgroup conferencing the DARPA Net (with TCP/IP and ethernet protocols) is inadequate. What is needed in addition is desktop video teleconferencing at standard voice telephone-line rates. Video teleconferencing is currently achieved today using expensive T1-class satellite links ($650/hour for 1544 Kbits/sec broadcast-quality bandwidth) and expensive codec devices to digitize the analog video signals at each end (est. $70K each). The possibility of using leased ISDN high-quality telephone lines (128 Kbits/sec) with ($2K codecs and Data Terminal Adaptors [DTAs] under a synchronous v.35 protocol) will revolutionize telecommunications in the next few years and radically change our transportation/communication tradeoffs. SMDS (Switched MultiMegabit Data Services) and Broadband ISDN (2.5 GB/sec using fiber optic cable) are expected by 1995.

The software needed to facilitate these kinds of multiperson desktop teleconferencing-in-a-window environments doesn't exist yet. BBN Slate is a close approximation in so far as multimedia e-mail capabilities are concerned, but it runs only on UNIX machines at present. A nice feature of Slate is its ability to place automatically a digitized voice-annotation clip under an icon embedded at an appropriate point in the text of a larger document. So, for example, if the recipient of a Slate document sees such an icon in the text and clicks on it, he may see a pop-up stating that "so and so" recorded 12 seconds of speech at "such-and-such" a time on "this-or-that" date. "Do you want to hear it?" Another click and he will then hear the sound clip through a speaker attached to his machine. This method of intermixing voicemail narration into the text of document can enormously speed productivity for busy decision-makers.

Large Data Bases

FoxPro in the PC environment or Sybase, Oracle, or Informix in the UNIX environment are good choices for developing data base applications. Distributed data bases over several network servers will become commonplace in the future. Compuserve and other information retrieval services are just starting a major expansion, as home PCs become more popular. Rumba from Wall Data, Extra! from Attachmate Corp., and Select from DCA are macro retrieval packages to facilitate PC to Mainframe database communications. Data integrity (UPS) and security are important issues. DES encryption boards from Centel, Inc. and SmartID cards for user password authentication from Security Dynamics address some of these issues.

Sound

Speech and music will both play an increasingly important role in software development in the future. Reasonably intelligible speech synthesis (text to speech) by hardware (SoundBlaster Board) or Software (Monolog 2.0) can be incorporated inexpensively into any software application to provide warnings or error messages or for whatever purpose. Speech input from the user, however, is much more difficult. The Introvoice VI board, for example, at a price of over $1000 is vocabulary limited (500-1000 words or phrases) to isolated

(noncontinuous) speech, speaker dependent, limited by having to endure an extensive training session, and is still unreliable in a noisy environment. Nevertheless, for certain applications it is quite respectable. Sphinx from Carnegie-Mellon University is a step toward reliable, continuous speech, but is still experimental and not yet incorporated into a commercial product. Music and the mixing of multichannel audio sources (drag-and-drop icons of sound effects into the middle of a production) is extremely difficult to do well. This challenge frequently distinguishes a high-quality production from an off-shore or low-budget production even in professional environments where the video portion is impeccable. Achieving dolby surroundsound quality in amateur productions is still some years away. Achieving the acoustics of an anechoic chamber, an intimate room, a conference room, a concert hall, a cathedral, a football field, etc. should someday be done at the click of a mouse.

Integration and Interoperability

DDE (Dynamic Data Exchange) or Hot Links between multiple applications running within the same window environment or across multiple platforms through a network is one of the most challenging aspects of work group conferencing. Remember that it was not too long ago that even tape formats were incompatible across vendor boundaries because open systems had not yet been advocated by the leaders of the industry. Standards were slowly incorporated through the diligence of government sponsors and disinterested professional societies. Just as all electrical appliances must be compatible with certain power distribution parameters (frequency, voltage, etc.) within national borders but are usually incompatible across continental barriers where different arbitrary standards were adopted, all computing cycles should also be "vanilla" flavored. Obviously, there are many more degrees of freedom in the parameterization of computing interfaces than there are in electric power distribution, but we are correspondingly more immature in our development of computing standards and this should be ameliorated in the coming decade. Differing man-machine interface standards for something as seemingly simple as "mouse clicking" in different applications on the same platform have been known to result in "acute brain meltdown" for some of our users. In our opinion, the recent trend toward consortia among industry leaders will facilitate this process of standardization and open computing.

The EASL (End user Applications Software development Laboratory) Laboratory at JPL (Figure 1) was created to test the interoperability of software across a number of the popular hardware platforms (DEC, HP, Sun, IBM, and Apple). EASL is less interested in benchmarking algorithms on different machines to measure performance than it is in verifying the operability of commercial software applications within the JPL network environment before major software acquisitions are finalized.

PRODUCTIVITY ENHANCEMENT

As a single testimonial, Micromedia Development Corp. is a software developer and audio/video supplier located in Vancouver, British Columbia that now uses video tutorials to replace written manuals for their systems. "We used to be inundated with technical support calls," said Roland Haynes. "With video tutorials, our calls have been reduced by 80 percent because we can compress a 400-to-500 page manual into a 30-minute video, which is even more effective." According to Haynes, "I would estimate that we have saved about $35,000 this year because three people can now support a work load that previously required eight or nine people."

SUMMARY

In today's market-driven world, multimillion-dollar 30-second TV commercials and longer video productions are crafted by teams of highly paid professionals. This price of admission is, of course, a major "barrier to entry" for ordinary folks. In the future, through the use of low-cost hypermedia-presentation technology, the production of high-quality targeted video presentations will no longer be limited by the thickness of one's pocketbook; relatively speaking, your capacity to communicate, to educate, and to persuade will only be limited by your imagination. In our view, as teams of software developers using this technology begin to interactively debug large complex systems, independently of geographic boundaries, our national productivity will be significantly enhanced.

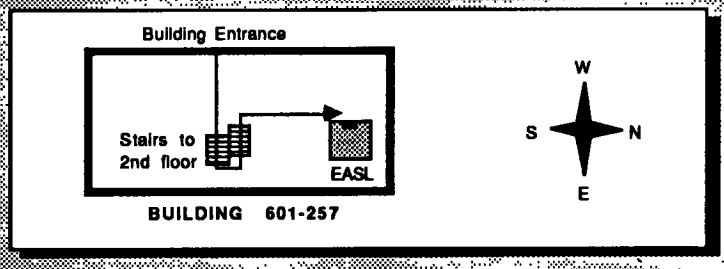
ACKNOWLEDGMENTS

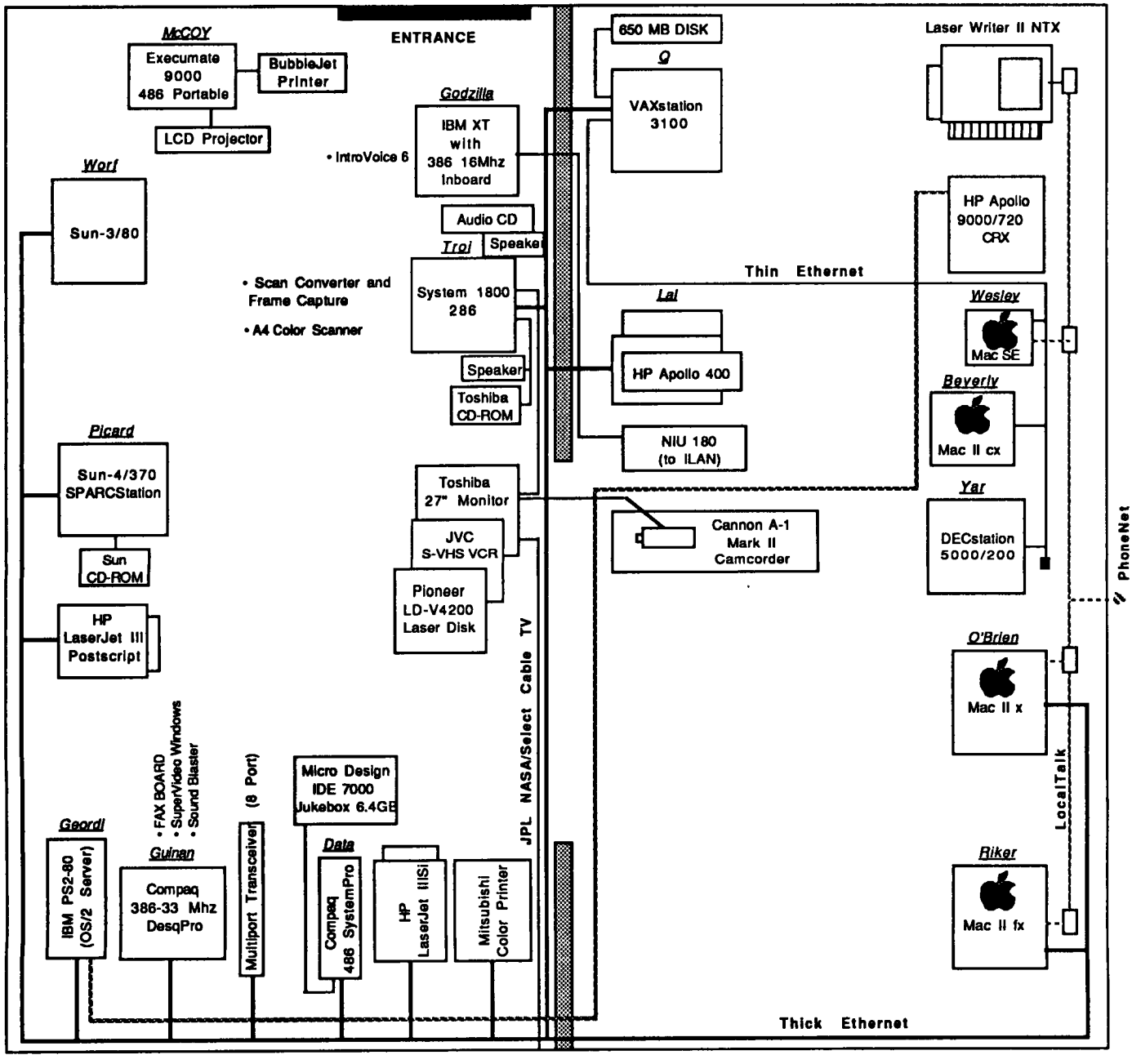
The work described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

JPL Section 373

End-User Applications Software Laboratory (EASL)





Last Update: 9-24-91

ADVANCED MANUFACTURING

(Session E1/Room A1)

Thursday December 5, 1991

- **Intelligent Robotic System with Dual-Arm Dexterous Coordination and Real-Time Vision**
- **Neural Network Software for Distortion-Invariant Object Recognition**
- **Constraint-Based Scheduling**
- **COMPASS: A General-Purpose Computer-Aided Scheduling Tool**

PRECEDING PAGE BLANK NOT FILMED

N92-22712

A COST-EFFECTIVE INTELLIGENT ROBOTIC SYSTEM WITH DUAL-ARM DEXTEROUS COORDINATION AND REAL-TIME VISION

Neville I. Marzwell
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

Alexander Y. K. Chen*
Scientific Research Associates, Inc.
P.O. Box 1058
50 Nye Road
Glastonbury, CT 06033

ABSTRACT

Dexterous coordination of manipulators based on the use of redundant degrees of freedom, multiple sensors, and built-in robot intelligence represents a critical breakthrough in development of advanced manufacturing technology. A cost-effective approach for achieving this new generation of robotics has been made possible by the unprecedented growth of the latest microcomputer and network systems. The resulting flexible automation offers the opportunity to improve the product quality, increase the reliability of the manufacturing process, and augment the production procedures for optimizing the utilization of the robotic system. Moreover, the Advanced Robotic System is modular in design and can be upgraded by closely following technological advancements as they occur in various fields. This approach to manufacturing automation enhances the financial justification and ensures the long-term profitability and most efficient implementation of robotic technology. The new system also addresses a broad spectrum of manufacturing demands and has the potential to address both complex jobs as well as highly labor-intensive tasks.

The Advanced Robotic System prototype employs the Decomposed Optimization Technique in spatial planning. This technique is implemented to the framework of the Sensor-Actuator Network to establish the general-purpose geometric reasoning system. The developed computer system is a multiple microcomputer network system, which provides the architecture for executing the modular network computing algorithms. The knowledge-based approach used in both the robot vision subsystem and the manipulation control subsystems results in the real-time image processing vision-based capability. The vision-based task environment analysis capability and the responsive motion capability are under the command of the local intelligence centers. An array of ultrasonic, proximity and optoelectronic sensors is used for path planning.

The Advanced Robotic System currently has 18 degrees of freedom made up by two articulated arms, one movable robot head and two CCD cameras for producing the stereoscopic views, an articulated cylindrical-type lower body, and an optional mobile base. A functional prototype will be demonstrated.

INTRODUCTION

Robotics is a science that analyzes the motion behavior of multiple correlated entities. Robotic engineering is a new engineering discipline that utilizes the knowledge of robotics to achieve those objectives that were previously unattainable unless performed by trained human beings. A generic robot would integrate the actuation of articulated motion, sensor processing, and machine perception into one system.

The application of robots has virtually no limitation, and the major obstacle of robot implementation today is the lack of proper understanding of robotics. In order to achieve what we need most, the robot applications can be categorized into three different levels: utilizing the robot as a function-specific tool, employing the robot as an

* Dr. Chen is currently with United Technologies Research Center, East Hartford, CT 06108.

PRECEDING PAGE BLANK NOT FILMED

intelligent machine, or constructing the robot as a self-contained artificial creature. Some examples of each level will be introduced to demonstrate the potential usages of modern robotic systems.

The Robot as a Tool

One major criterion for differentiating robots from conventional electromechanical systems is onboard machine intelligence. Current microcomputer technology makes it possible to embed certain autonomous functions in the articulated mechatronic infrastructure. Although there are many intrinsic limitations, such as memory size, processing speed, and hardware constraints, the resulting robotic systems are generally capable of performing a set of specific functions with superior quality. Typical examples are advanced manufacturing processes, microscale devices fabrication, and precision laboratory automation.

The Vital Component of Flexible Manufacturing Systems (FMS). Current manufacturing systems are experiencing a new challenge when the traditional approach is no longer capable of matching the dynamic environment of market demand. As global competition prevails in every sector of the industrial world, the required manufacturing system has to be cost effective, schedule responsive, quality sensitive and long-term operationally stable. As the availability of information and self-education tools becomes reality, long-hour routine jobs are no longer suitable for human labors. Instead, different robotic tools will gradually fill in this gap to constitute the vital component of the flexible manufacturing systems. It is expected that a completely automated manufacturing system would provide reprogrammability, which is essential to accommodate a wide range of production demands, maximize the utilization rate of the available equipment, and maintain uniform product quality.

Furthermore, the incorporation of robotic tools would make the FMS a feasible production solution for those companies of medium or small size as well. Since those labor-intensive operations are minimized, the frequency of repair and maintenance is substantially reduced, and the supporting manpower is consolidated; consequently, the competitiveness of the manufacturing industry will be increased significantly. Once the technology upgrade procedure is properly established, the modern manufacturing technology realized by various function-specific robotic systems will help our country to regain the leading position in every industrial sector.

Application-Specific Integrated Devices (ASID). Several major sectors of our society are encountering severe difficulty in recruiting human workers, e.g. hospitals, schools, law enforcement in major metropolitan areas, fire department and sanitary operations, etc. This phenomenon is significantly affecting our living conditions, and it does not appear that it will improve automatically. There may be some associated social reasons that should be reckoned with. However, it is conceivable that with the help of modern technology, a series of application-specific integrated devices (ASIDs) may satisfy these pressing needs and reverse the trend of downgrading our living standard. Robotic engineering can serve as the major ingredient for integrating the existing electromechanical components with application-specific functionalities to develop the required ASIDs. For instance, ASIDs for nursing can immediately help patients to perform some basic functions wherever and whenever human nurses are not available; ASIDs for day care can assist babysitters by taking care of some routine jobs when they are occupied. All kinds of ASID products can not only utilize currently available technology well to improve our lives, but also will relieve social tension contributed by social restructuring and prolonged life expectancy.

The Robot as an Intelligent Machine

When robot intelligence has developed enough to become mature in dealing with heterogeneous matters, more responsibility can be assigned to robotic systems. Then robots will behave as intelligent machines that can handle a group of related tasks without human intervention. The principle of robotics implicitly indicates a progressive path along which robots are not only application-specific tools, but also can be intelligent machines. The utilization of information, knowledge, and experience has been well formulated such that the capability of handling a group of related jobs can be downloaded to the robot and constitutes a machine with confined intelligence. Two examples are introduced below to point out some potential near-term objectives in developing the robotic system as an intelligent machine.

A Reliable Means of Optimizing Human Resources. The transfer of low-level work from well-developed countries to developing countries is an apparent trend that is generally accepted by the majority of the world. As this transfer progresses, in the near future there will be a lot of low-level jobs that human workers will no longer be content with. Unless the evolution of human history is reversed, this trend is inevitable. By that time, it would be unquestionable that robots would be the most reliable source for fulfilling the need for low-level work. In particular,

labor-intensive, hazardous, repetitive jobs, such as are found in farming, mining, fishing, and security monitoring, are experiencing an urgent need to provide a front-end work force that is highly reliable and yet flexible enough to adapt to variable environments, cost-effective, and also quality sensitive. As long as the threshold of economics is overcome, the employment of robots as the intelligent front-end machines would be commonly acceptable.

An Indispensable Helper for the Physically Challenged. Mobility is one of the major qualities that we in our daily living cannot do without. For physically challenged people, however, maintaining the basic level of mobility can be unthinkably difficult without the assistance of a human companion. Robotic systems have been considered to be the best solution to compensate for disability if the incorporated robotic device is intelligent enough to deal with unknown situations and to communicate directly with the neural system of the host. In examining the latest progress in medical science as well as in robotics, it is apparent that future robotic ASIDs can provide indispensable help for the physically challenged. Moreover, robotic ASIDs for medical applications can not only assist human beings externally, but also can be implanted into the human body to provide critical functions internally.

The Robot as a Man-Made Creature

Once the robotic system can perform intelligent functions with an onboard knowledge base, the next stage will be the development of self-learning capability. In order to build an autonomous system, it is essential that the robot be able to self-adjust both the database and the rule base automatically. In other words, not only could the system variable values be updated and the modeling parameters adjusted, but the rules of formulating the internal representation of the controlled environment could also be modified. With layers of inferencing, a certain self-organizing capacity would then be established to provide a robust decision-making procedure that would be independent of the employed system models. Therefore, the resulting autonomous robot can be regarded as a man-made creature that is self-sufficient for accomplishing the assigned mission. It is expected that as the research and development of autonomous robot intelligence approaches maturity, related biotechnology will also reach the stage where artificially grown biocells can directly communicate with the electronic subsystems. Then, with the merging technology of biorobotics, the issue of self-propagation can be addressed as well. The following two examples are typical cases of utilizing robots as man-made creatures.

Future Explorers of the Unknown Universe. The exploration of the entire universe is considered as the last frontier of human knowledge expansion. However, existing knowledge about human physical systems severely limits the feasibility of space exploration with human presence. Since our imaginative capability and brainpower exceed our physical limitations, it would be more appropriate to send some autonomous robots into the galaxy to help us ω explore the unreachable universe before the development of technology for overcoming our physical barriers. It is also likely that the results of autonomous robot explorations will accelerate the growth of human knowledge in terms of better understanding of the three-dimensional physical world.

Settlers in New Territory Unfit for Mankind. The other planets, of which we already have some preliminary knowledge, more often than not indicate a specific physical condition that is not suitable for human living. However, there is plenty of information that needs to be collected and investigated and for which physical contact is necessary. Instead of spending an enormous amount in resources to generate a tiny compartment to accommodate a human's physical limitations, it would be more justifiable to utilize autonomous robots or semiautonomous telerobots as the settlers on those planets as the first stage of exploration. It is believed that the research and development of human space exploration should be preceded by that of space robot exploration.

Robotics as an Application Technology. Within this NASA-funded SBIR project, six technical advancements have been accomplished to demonstrate the developed robotic engineering capability. If robotics is regarded as an application technology, the developed technical advancements are readily available for various industrial implementations.

ROBOTIC-SYSTEM MODULARIZATION

The idea that a reconfigurable, modular robot design can best utilize the advantages of robotics technology as it evolves is a cost-effective one [1,2]. In addition to the analytical development of modular architecture of robotic systems, there are several technical issues that need to be addressed before the successful implementation of modular robot design can take place. First of all, it is essential to incorporate the actuator directly at each active joint; second, the connections between adjacent modules have to be standardized such that the interchange capability can be

maintained; also, intelligence localization and distributed control are necessary to maintain the modular functionality [3]. In this program, the modular design of the electromechanical system is capable of decomposing the robotic system into five modules. Each module has distinct functionalities. The corresponding spatial planning, kinematic planning, and motion execution are also considered and designed in a modular fashion, which involves analytical framework development, algorithmic construction, software coding, and system communication protocols. Eventually, a joint-oriented modular mechatronic system will be available to satisfy various demands from simple routine jobs to high-difficulty tasks. A proposed joint-oriented intercoordinate system is shown in Figure 1.

UTILIZING REDUNDANT DEGREES OF FREEDOM WITH AISP

One of the advantages of utilizing the Artificial Intelligence for Spatial Planning (AISP) technique is to carry out the online task planning by systematically maneuvering the multiple degrees of freedom of the redundant robots [4]. Since there are many applications that require the redundant degrees of freedom of the robotic system to perform articulated manipulations, the restriction to a maximum of six or seven degrees of freedom in existing robotic systems substantially constrains the applicability of robots. One of the main reasons for accepting this constraint is the lack of implementable techniques to control redundant robots. The proposed AISP implementation, which utilizes the Decomposed Optimal Transition (DOT) technique for solving the generalized reachability problem [5], presents a major improvement in expanding the domain of robotics technology implementation.

In a three-dimensional computer graphics simulation, the Advanced Robotic System had an assignment to reach two designated locations, one with specified position and the other with specified orientation. The initial scenario is shown in Fig. 2(a), where the two triangles stand for the given targets. The final scenario is presented in Fig. 2(b), with the top view in Fig. 2(c). With the help of AISP, the robot would be able to reach the assigned target locations without touching the cylindrical obstacle between the robot and the targets.

STEREOSCOPIC REAL-TIME ROBOT VISION

The development of a novel robot vision system falls under the category of advanced sensor technology. The lower level of the image processing produces the edge representation of two asynchronous views as the result of a stereoscopic arrangement of two CCD cameras. The arrangement is shown in Fig. 3. The medium level of image processing performs the basic logic for scene analysis. Both the object identification and the motion analysis are addressed. The results from a relational database are then used for processing at a higher level. By communicating with the corresponding knowledge base, some fundamental geometric reasoning capability can be established to extract the necessary environmental information to support various tasks of robot manipulation. The functional diagram of the developed robot vision system is shown in Fig. 4, and the details can be found in [6].

RESPONSIVE CONTROL IN THE SENSOR-ACTUATOR NETWORK

The sensor-actuator network (SANE) represents the first layer of robot behavior. In order to react to some critical situations as soon as they are sensed, a set of responsive control rules is needed that has the highest priority and will react to the unexpected causes and minimize the potential system damage. Responsive control would be able to override the standard actuation sequence and execute the appropriate emergency action. The sensor-driven interrupts would be terminated only when the alarming situation no longer existed or the related subsystem was (or subsystems were) shut down. The responsive control not only is the intricate part of the system's contingency plan, but also acts as the basic machine intelligence to prevent the user-specified program from making some trivial errors. The development of the responsive control system for the SANE of the Advanced Robotic System demonstrates the fundamental concept of this approach. The incorporation of a neuro-fuzzy data fusion process will be the next step to be pursued in the future research and development of autonomous robots and intelligent telerobotic systems.

THE DYNAMIC KNOWLEDGE EVOLUTION PROCESS

The dynamic knowledge evolution (DKE) process is an innovative approach in developing knowledge systems. For robot intelligence development, the DKE process is utilized to establish the basic geometric reasoning and the modeling and analysis of the environment encountered. In addition to having the common perception of expert systems, the knowledge systems with the DKE process would accept online knowledge base updates, modify rule-based inferencing to optimize the processing speed, and generate consistency rectification with analogy learning and probabilistic reasoning (in particular, Bayesian inferencing [7]). Due to the associated dynamic behavior, the theoretical foundation of DKE has a close relationship with the neural network theory explained in [8]. The

resulting FULOSONN (FUzzy LOgic and Self-Organized Neural Network) technique is a new technique in knowledge engineering. Figure 5 shows the functional diagram of the developed control system, and the intelligence flow is depicted in Fig. 6.

MuMicS ARCHITECTURE FOR MACHINE AUTONOMY

The advantage of integrating multiple microprocessors has gradually been appreciated for various applications, such as image processing, fault-tolerant systems, or large-scale real-time computations. Instead of closely coupled microprocessors, a new architecture incorporating multiple microcomputers is introduced. The advantages of MuMicS are the high reconfigurability (which is especially suitable for modular system design), the fault tolerance capability, and multiple asynchronous (or synchronous) tasking. Due to the rapid growth of computer technology, the MuMicS architecture design not only is functionally desirable but also provides a cost-effective solution for numerous applications. The Advanced Robotic System utilizes five microcomputer systems to constitute the control, command, communication and intelligence (C3I) system. One of them is the global intelligence center (GIC), which acts as the brain of the Advanced Robotic System. The remaining four are local intelligence centers (LIC): one LIC controls the right arm, one LIC controls the left arm, one LIC controls the mobile head with vision subsystem, and the last LIC controls the lower body. The functionality of the GIC is shown in Fig. 7. Each center has different functionalities and is connected through a local area network (LAN) system. The completion of the Advanced Robotic System protoflight demonstrates the feasibility of using the proposed MuMicS architecture to develop advanced autonomous robots.

CONCLUSION

Emphasizing robotics as an application technology has been the major guideline for constructing the Advanced Robotic System. Due to the limited resources available, this program has been constrained in its attempts in research and prototyping a systematic approach to developing implementable robotics technology. Advances along six technology thrusts critical to the growth of robotic engineering as a new discipline have been accomplished. The prototype Advanced Robotic System developed in this effort and shown in Fig. 8 can be used as the basis of future robotic engineering development. It is firmly believed that the continuation of all six thrusts in the technological development is necessary to maintain the vital competitiveness of robotic engineering in the United States. It is sincerely hoped that our effort in this program will eventually cause others to regard robotics as a vital technology that can impact our progress and competitiveness.

ACKNOWLEDGMENTS

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This work was funded by a NASA/SBIR contract from the Jet Propulsion Laboratory.

REFERENCES

1. Tourasis, V.D., and Ang, M.H., Jr., "A modular architecture for inverse robot kinematics," *IEEE Trans. Robotics and Automation*, Vol. 5, No. 5, pp. 555-568, October 1989.

2. Schmitz, D., Hoffman, R., and Khosla, P.K., "CHIMERA: A real-time programming environment for manipulator control," Technical Report CMU-RI-88, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, September 1988.

3. Chen, A.Y.K., and Chen, E.Y.S., "Intelligent manipulation technique for multi-branch robotic system—quarterly report," SRA Report No. R90-900077-3, Scientific Research Associates, Inc., Glastonbury, Connecticut, September 1990.

4. Chen, A.Y.K., "AISP—A robot intelligence system in spatial manipulation," *Proceedings of ROBEXS '89, The Annual Workshop on Robotics and Expert Systems*, Palo Alto, California, August 2–4, 1989.

5. Chen, A.Y.K., "Decomposed optimization technique for two-dimensional N-link serial robot arm," Connecticut State SBIR Program Final Report, January 1989.

6. Chen, A.Y.K., and Chen, E.Y.S, "Intelligent vision process for robot manipulation," SPIE Symp. Intelligent Systems, Boston, Massachusetts, November 1990.

7. Kosko, B., *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Prentice Hall Publishing Co., Englewood Cliffs, New Jersey, 1991.

8. Box, G.P., and Tiao, G.C., *Bayesian Inference in Statistical Analysis*, Addison-Wesley Publishing Co., Reading, Massachusetts, 1973.

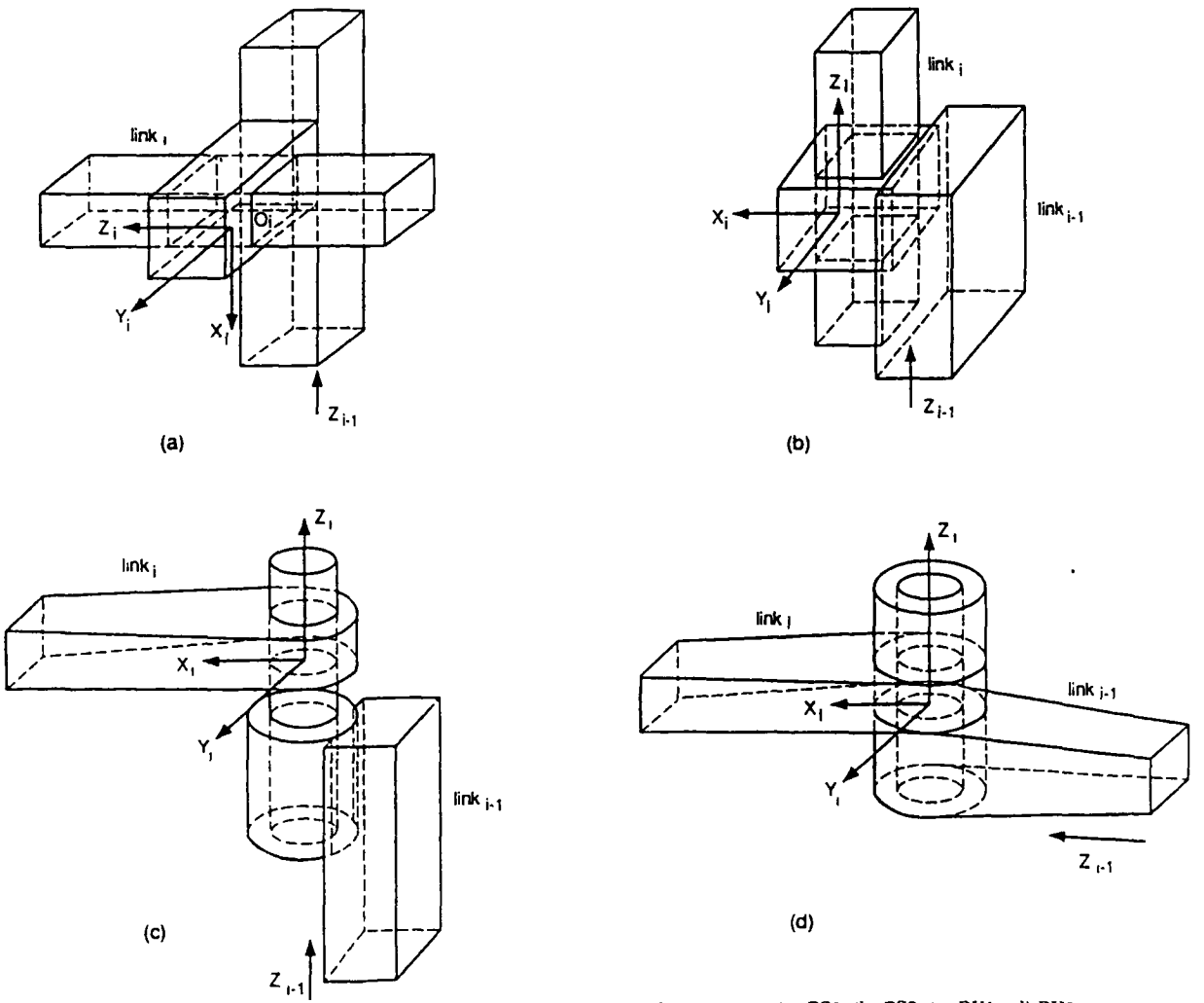


Figure 1. The joint-oriented convention: (a) PS1, (b) PS2, (c) RV1, (d) RV2.

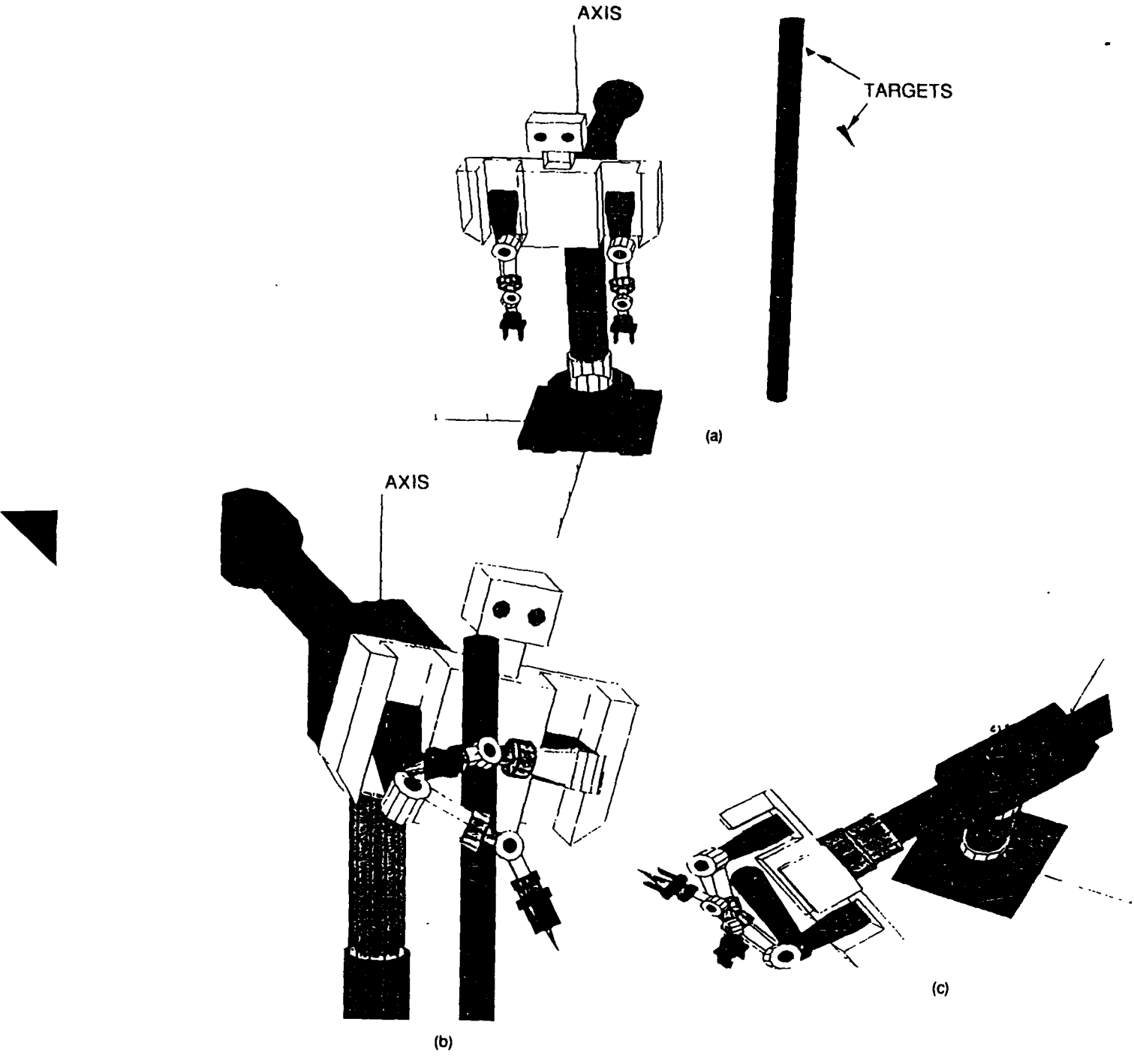


Figure 2. Three views of AISP with collision avoidance: (a) initial scenario, (b) final scenario, (c) top view.

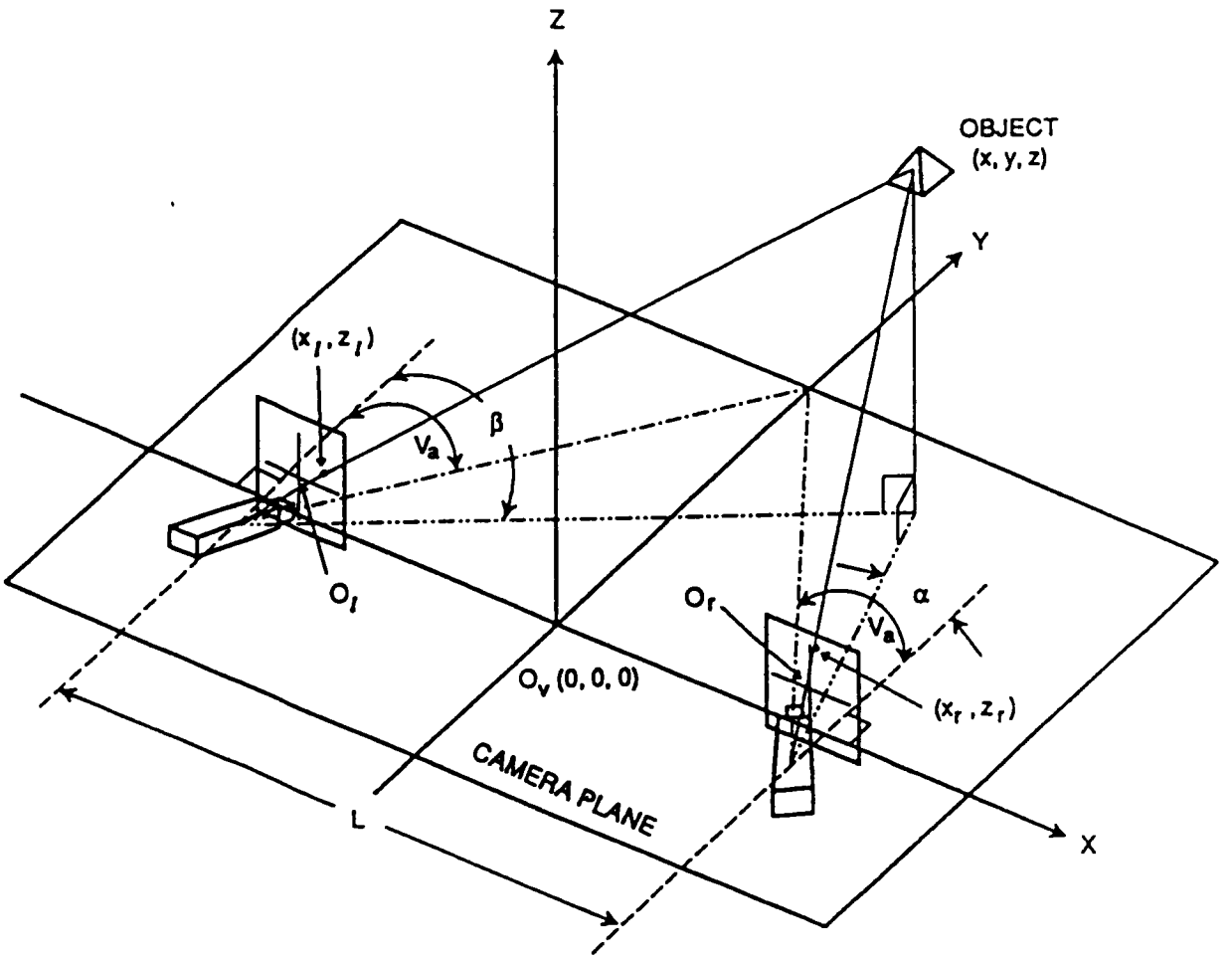


Figure 3. The three-dimensional positioning diagram in the developed stereoscopic viewing system.

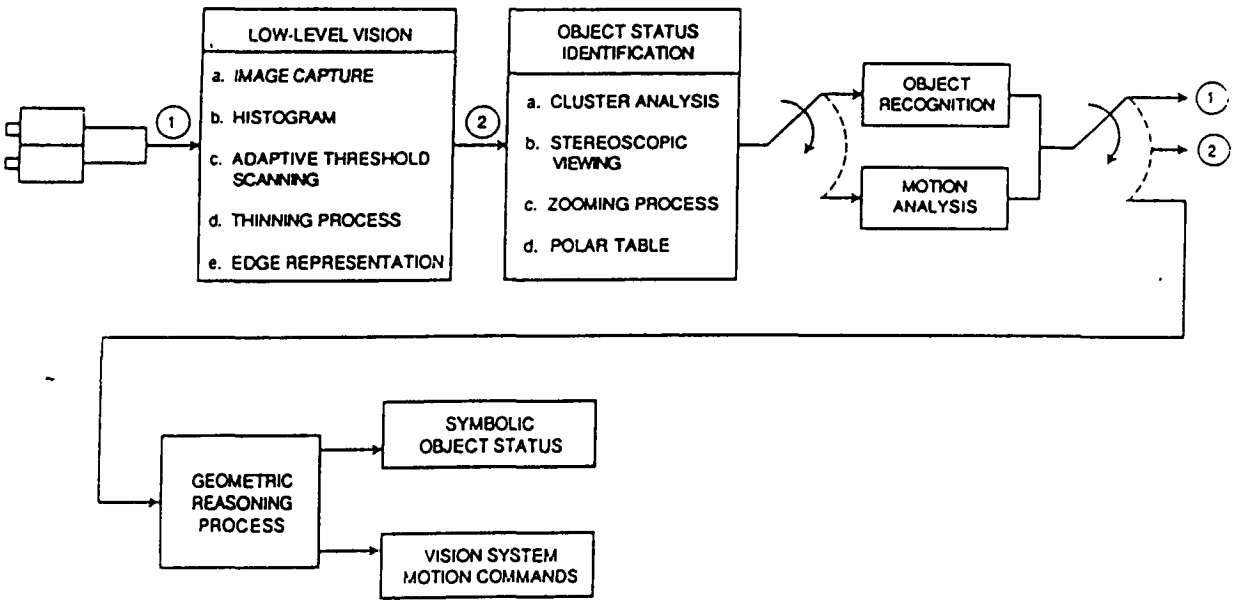


Figure 4. The functional diagram of the new robot vision system.

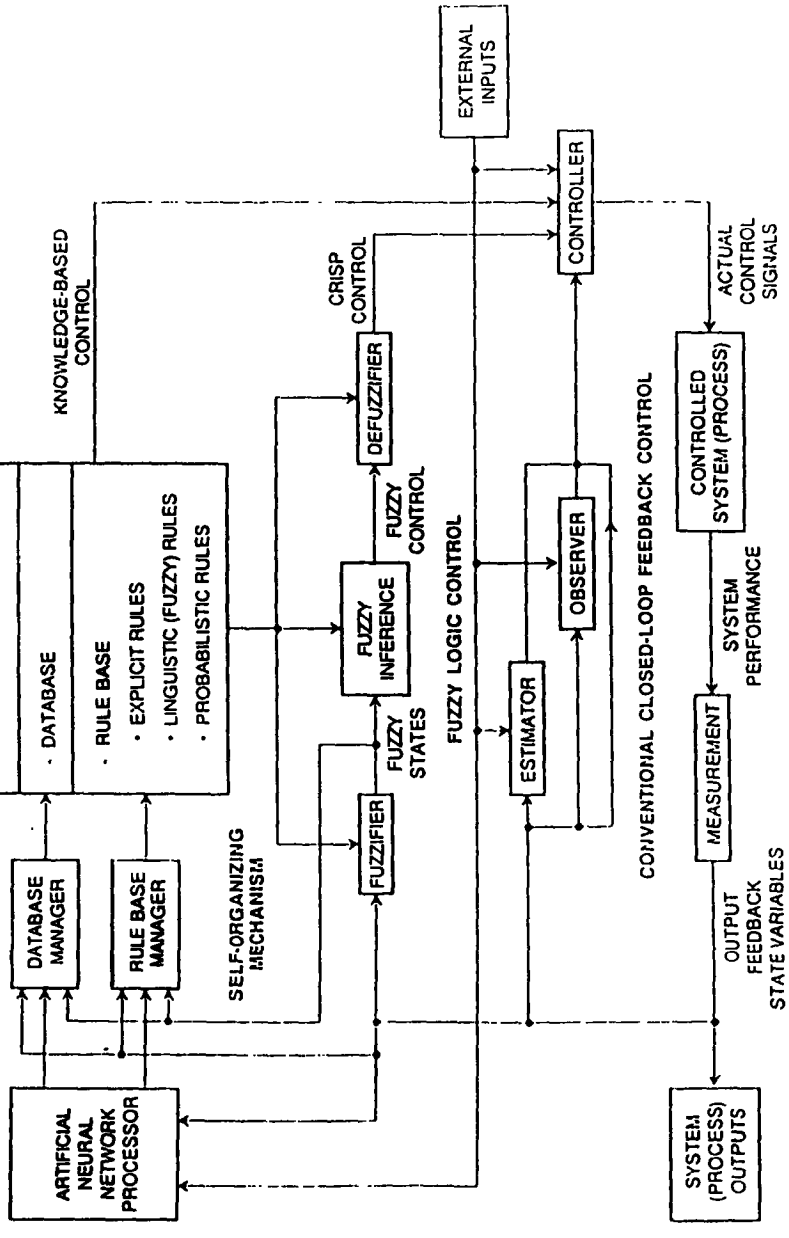


Figure 5. The functional diagram of modern control technology.

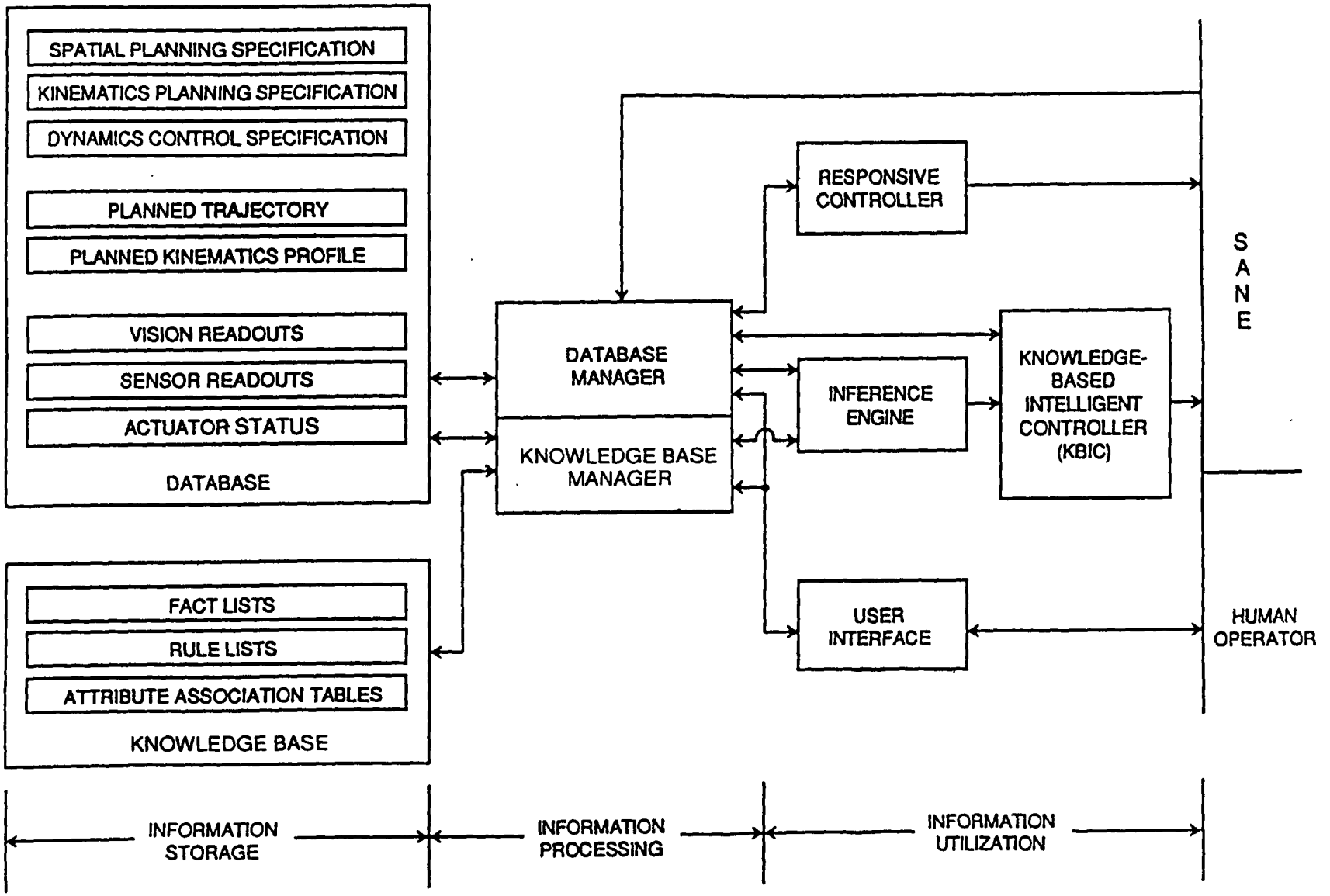


Figure 6. The functional diagram of KBIC with DKE processes.

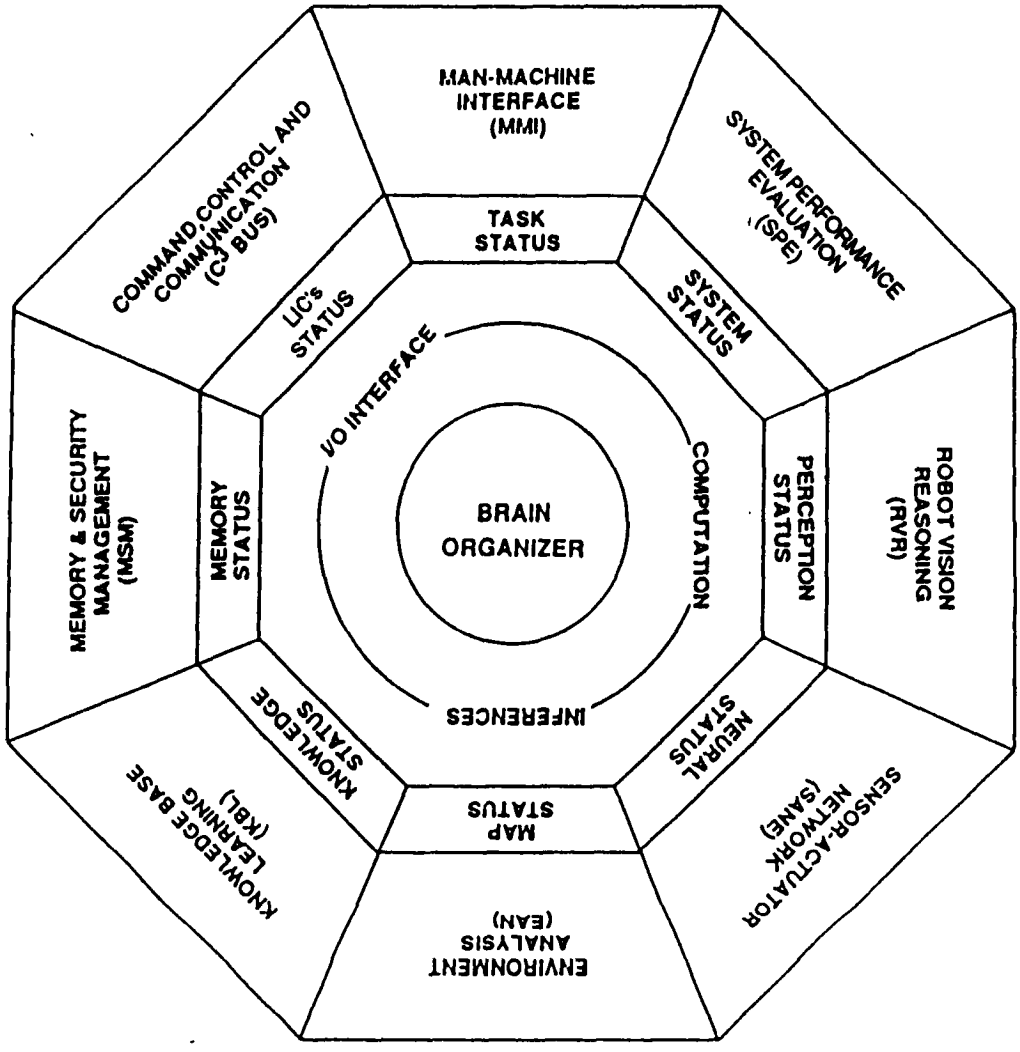


Figure 7 Functional diagram of SRAARS's brain

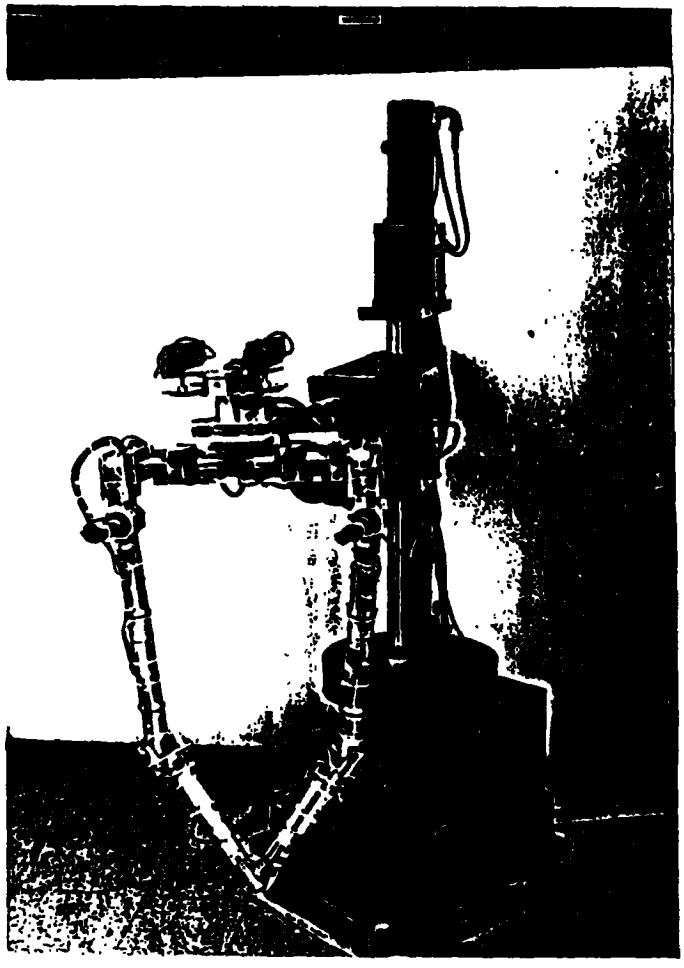


Figure 8. The advanced robotic system prototype.

N92-22713

HIGHER-ORDER NEURAL NETWORK SOFTWARE FOR DISTORTION INVARIANT OBJECT RECOGNITION

Max B. Reid and Lilly Spirkovska

NASA Ames Research Center
Mail Stop 244-4
Moffett Field, CA 94035

ABSTRACT

The state-of-the-art in pattern recognition for such applications as automatic target recognition and industrial robotic vision relies on digital image processing. Digital image processing for automatic pattern recognition is very computationally intensive, involving feature extraction performed via large matrix operations. Digital techniques for recognizing objects regardless of their position, scale, and angular orientation are even more computationally intensive and cannot run at real time. They also are not readily adaptive due to the long time required to compute the matrix equations in digital algorithms.

We present a higher-order neural network model and software which performs the complete *feature extraction - pattern classification* paradigm required for automatic pattern recognition. Using a third-order neural network, we demonstrate complete, 100% accurate invariance to distortions of scale, position, and in-plane rotation. In a higher-order neural network, feature extraction is built into the network, and does not have to be learned. Only the relatively simple classification step must be learned. This is key to achieving very rapid training. The training set is much smaller than with standard neural network software because the higher-order network only has to be shown one view of each object to be learned, not every possible view.

The software and graphical user interface run on any Sun workstation. We also present results of the use of the neural software in a autonomous robotic vision system. Such a system could have extensive application in robotic manufacturing.

I. INTRODUCTION

Neural networks have been applied to various domains including speech recognition, trend analysis and forecasting, process monitoring, robot control, and object recognition. We present work in the position, scale, and rotation invariant (PSRI) object recognition domain. The objective in this domain is to recognize an object despite changes in the object's position in the input field, size, or in-plane orientation, as shown in Figure 1.

Pattern recognition may be viewed as a two part process of feature extraction followed by object classification[1-2]. First, a preliminary mapping from an image to a representation space is made, generally resulting in a significant degree of data reduction. A second mapping then operates on this reduced data to produce a classification or estimation in an interpretation space. Historically, these steps have required mathematical mappings operating directly on a detected image. However, digital image processing techniques are very computationally intensive, require extensive computer calculations, and have difficulty handling full in-plane distortion invariance.

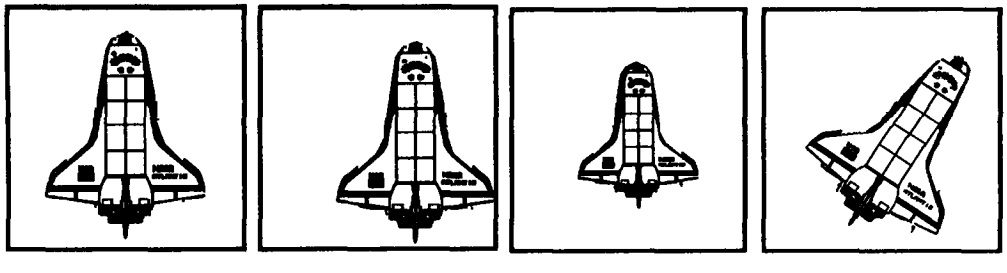


Figure 1: PSRI object recognition. In the PSRI (position, scale, and rotation invariant) object recognition domain, all four of these objects would be classified as a single object. Three distortions of the prototype in (a) are shown. The object in (b) is a translated view, (c) is scaled, and (d) is rotated in-plane.

In this paper we discuss higher order neural networks as implementations of the complete pattern recognition operation. Higher-order neural networks can be designed to implement the extraction of simple but effective features suitable for in-plane distortion invariance. Known geometric relationships are exploited and the desired invariances are built directly into the architecture of the network. Building such domain specific knowledge into the network's architecture results in a network which is pre-trained and does not need to learn invariance to distortions. For each new set of training objects, a HONN only needs to learn to distinguish between one view of each training object; it does not need to be trained on all distorted views. Therefore, training time is reduced significantly from that typically required for other neural models. Moreover, 100% recognition accuracy is guaranteed for noise-free images characterized by the built-in distortions.

We explain how known relationships can be exploited and desired invariances built into the architecture of higher-order neural networks, discuss some limitations of HONNs and how to overcome them, present simulation results demonstrating the usefulness of HONNs with practical object recognition problems, discuss the performance of HONNs with noisy test data, and present laboratory results of using a HONN to control a robot performing a manufacturing task.

II. HIGHER-ORDER NEURAL NETWORKS

The output of a node, denoted by y_i for node i, in a general higher-order neural network is given by

$$y_i = \Theta\left(\Sigma_j w_{ij} x_j + \Sigma_j \Sigma_k w_{ijk} x_j x_k + \Sigma_j \Sigma_k \Sigma_l w_{ijkl} x_j x_k x_l + \ldots\right) \quad (1)$$

where $\Theta(f)$ is a nonlinear threshold function such as the hard limiting transfer function given by

$$\begin{aligned} y_i &= 1, \text{ if } f > 0, \\ y_i &= 0, \text{ otherwise,} \end{aligned} \quad (2)$$

the x's are the excitation values of the input nodes, and the interconnection matrix elements, w, determine the weight that each input is given in the summation. Using information about relationships expected between the input nodes under various distortions, the interconnection weights can be constrained such that invariance to given distortions is built directly into the network architecture [3-7].

For instance, consider a second-order network as illustrated in Figure 2. In a second-order network, the inputs are first combined in pairs and then the output is determined from a weighted sum of these products. The output for a strictly second-order network is given by the function

$$y_i = \Theta\left(\Sigma_j \Sigma_k w_{ijk} x_j x_k\right). \quad (3)$$

Pattern recognition invariant to geometrical distortions in the object are achieved by constraining the values which the weights w_{ijk} are allowed to take on.

As an example, each pair of input pixels combined in a second-order network define a line with a certain slope. As shown in Figure 3, when an object is moved or scaled, the two points in the same relative positions within the

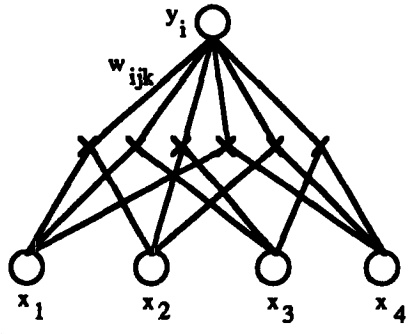


Figure 2: Second-order neural network. In a second-order neural network, the inputs are first combined in pairs (at X) and the output is determined from a weighted sum of these products.

object still form the endpoints of a line with the same slope. If all pairs of points which define the same slope are connected to the output node using the same weight, the network will be invariant to distortions in scale and translation. In particular, for two pairs of pixels (j, k) and (l, m), with coordinates (x_j, y_j), (x_k, y_k), (x_l, y_l), and (x_m, y_m) respectively, the weights are constrained according to

$$w_{ijk} = w_{ilm}, \quad \text{if } (y_k - y_j) / (x_k - x_j) = (y_m - y_l) / (x_m - x_l). \tag{4}$$

Alternatively, the pair of points combined in a second-order network may define a distance. As shown in Figure 4, when an object is moved or rotated within a plane, the distance between a pair of points in the same relative position on the object does not change. If all pairs of points which are separated by equal distances are connected to the output with the same weight, the network will be invariant to translation and in-plane rotation distortions. The weights for this set of invariances are constrained according to

$$w_{ijk} = w_{ilm}, \quad \text{if } \|d_{jk}\| = \|d_{lm}\|. \tag{5}$$

That is, the magnitude of the vector defined by pixels j and k (d_{jk}) is equal to the magnitude of the vector defined by pixels l and m (d_{lm}).

To achieve invariance to translation, scale, and in-plane rotation simultaneously, a third-order network can be used. The output for a strictly third-order network, is given by the function

$$y_i = \Theta \left(\Sigma_j \Sigma_k \Sigma_l w_{ijkl} x_j x_k x_l\right). \tag{6}$$

Each set of input pixel triplets forms a triangle with some included angles (α, β, γ), as shown in Figure 5. When an object is translated, scaled, or rotated in-plane, the three points in the same relative positions on the object still form the included angles (α, β, γ). In order to achieve invariance to all three distortions, all sets of triplets forming similar triangles are connected to the output with the same weight. That is, the weight for the triplet of inputs (j, k, l) is constrained to be a function of the associated included angles (α, β, γ) such that all elements of the alternating group on three elements (group A3) are equal

$$w_{ijkl} = w(i,\alpha,\beta,\gamma) = w(i,\beta,\gamma,\alpha) = w(i,\gamma,\alpha,\beta). \tag{7}$$

The fact that HONNs are capable of providing nonlinear separation using only a single input layer and a single output layer, with no hidden layer of nodes required, allows them to be trained using a simple rule of the form

$$\Delta w_{ijkl} = (t_i - y_i)\, x_j\, x_k\, x_l, \tag{8}$$

where the expected training output, t, the actual output, y, and the inputs, x, are all binary.

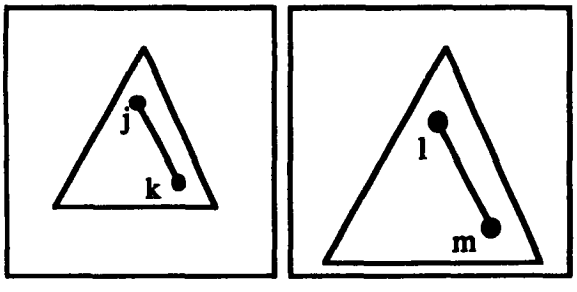


Figure 3: Translation and scale invariance in a second-order network. By constraining the network such that all pairs of points which define equal slopes use equal weights, translation and scale invariance are incorporated into a second-order neural network.

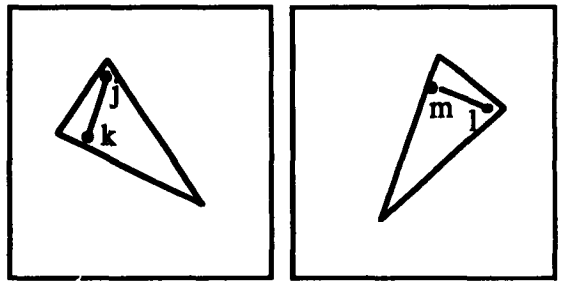


Figure 4: Translation and rotation invariance in a second-order network. By constraining the network such that all pairs of points which are equal distances away use equal weights, translation and rotation invariances are incorporated into a second-order network.

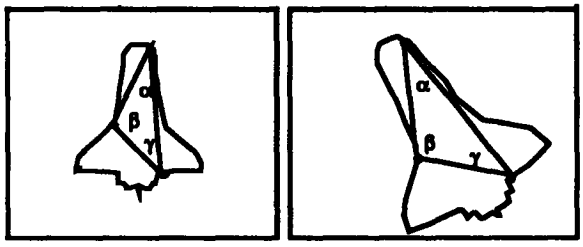


Figure 5: PSRI recognition with a third-order neural network. As long as all similar triangles are connected to the output with the same weight, a third-order network will be invariant to scale, in-plane rotation, and translation distortions.

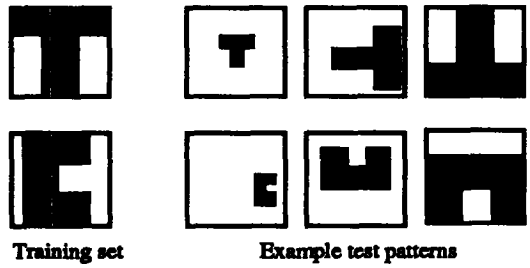


Figure 6: Training set and sample test patterns for distinguishing a "T" and a "C", invariant to translation, scale, and rotation.

The main advantage of building invariance to geometric distortions directly into the architecture of the network is that the network is forced to treat all distorted views of an object as the same object. Distortion invariance is achieved before any input vectors are presented to the network. Thus, the network needs to learn to distinguish between just one view of each object, not numerous distorted views, which leads to rapid convergence.

Software Results: Fully-connected Networks

We developed third-order network software using a Sun 3/60 workstation, where the third-order network was designed for scale, translation, and in-plane rotation invariance in a 9x9 pixel input field, giving 81 input nodes. The network had just one output node and one input layer. To build in invariance to distortions in scale, translation, and in-plane rotation, the weights were constrained according to Eq. (7) and the network was trained using the rule in Eq. (8).

The network was trained on just one view of each of the objects it was required to learn. In particular, we trained the network on the T/C recognition problem. As explained in Rumelhart [8], in the T/C problem, both objects are constructed of 5 squares, as illustrated in Figure 6, and the problem is to discriminate between them independent of translation or 90 degree rotations. In our work, the network was also required to distinguish between the objects invariant to distortions in scale.

The network learned to distinguish between all distorted views of a "T" and a "C" after just 10 passes through the training set, requiring less than 60 seconds on a Sun 3/60. The network was trained on just one view of a "T" and one view of a "C", as shown in Figure 6. Nevertheless, because the invariances are built into the architecture of the network, it was able to distinguish between the two characters regardless of their position in the input field, 90 degree rotations, or changes in size over a factor of three. In principle, recognition is invariant for any rotation angle, given sufficient resolution to draw the objects accurately.

III. EXPANDING TO PRACTICAL IMAGE SIZES

The advantages of HONNs stem from the fact that known relationships are incorporated directly into the architecture of the network. The network weights are constrained by this domain specific knowledge. Thus, fewer training passes and a smaller training set are necessary to learn to distinguish between the training objects.

The assumption behind incorporating specific knowledge into a network is that the weight values determined by the learning process result in the same output for one view of an object and a distorted view of the same object. Specifically, in our work, we assumed that the relationship expressed by Eq. (7), that all similar triangles have the same weight, constrained the network sufficiently so that an object and a distorted view of the same object would produce the same output. Using this relationship, we demonstrated that a third-order network can achieve simultaneous invariance to translation, in-plane rotation, and scale on the T/C recognition problem in a 9x9 pixel input field. Unfortunately, due to the finite resolution of actual images [7], Eq. (7) constrains the network adequately only in this limited domain but not when using a more general set of objects or a larger input field. Invariance to object scale changes can be lost when using larger image field sizes.

Problems arising from finite image resolution can be largely overcome by using edge-only images, as shown in Figure 7, and by restricting the resolution to which the angles α, β, and γ are calculated. We have shown that for a 36x36 pixel input field, angles need to be rounded to the nearest 20° in order for test objects to be recognized when scaled down to 50% of the training image size. As the input field is increased to 80x80 pixels, the angle resolution can be increased to the nearest 10°. Further increasing the input field resolution to 127x127 pixels allows the angle resolution to increase to 5°. Thus, with larger input fields, both the image resolution and the resolution to which α, β, and γ are calculated can be increased.

A greater constraint on increasing the size of images which can be evaluated using HONNs is the amount of storage required to implement the network. A network with M inputs and one output using only rth order terms requires M-choose-r interconnections. For large M, this number, which is on the order of M^r, is clearly excessive, as some storage must be used to associate each triplet of pixels with a set of included angles. In an NxN pixel input field, combinations of three pixels can be chosen in N^2-choose-3 ways. Thus, for a 9x9 pixel input field, the number of possible triplet combinations is 81-choose-3 or 85,320. Increasing the resolution to 128x128 pixels increases the number of possible interconnections to 128^2-choose-3 or 7.3×10^{11}, a number too great to store on most machines. On our Sun 3/60 with 30 MB of swap space, we can store a maximum of 5.6 million (integer) interconnections, limiting the input field size for fully connected third-order networks to 18x18 pixels. Furthermore, this number of interconnections ($\sim 10^{12}$) is far too large to allow a parallel implementation in any hardware technology that will be commonly available in the foreseeable future.

A coarse coding algorithm [7,9] can be used to permit an input field size practical for object recognition problems. The coarse coding algorithm involves overlaying fields of coarser pixels in order to represent an input field composed of smaller pixels, as shown in Figure 8. Figure 8a shows an input field of size 10x10 pixels. In Figure 9b, we show two offset but overlapping fields, each of size 5x5 "coarse" pixels. In this case, each coarse field is composed of pixels which are twice as large (in both dimensions) as in Figure 8b. To reference an input pixel using the two coarse fields requires two sets of coordinates. For instance, pixel (x=7, y=6) on the original image would be referenced as the set of coarse pixels ((x=D, y=C) & (x=III, y=III)), assuming a coordinate system of (A, B, C, D, E) for coarse field one and (I, II, III, IV, V) for coarse field two. This is a one-to-one transformation. That is, each pixel on the original image can be represented by a unique set of coarse pixels.

This transformation of an image to a set of smaller images can be used to greatly increase the resolution possible in a higher-order neural network. For example, a fully connected third-order network for a 10x10 pixel input field requires 10^2-choose-3 or 161,700 interconnections. Using 2 fields of 5x5 coarse pixels requires just 5^2-choose-3 or 2300 interconnections, accessed once for each field. The number of required interconnections is reduced by a factor of ~70. For a larger input field, the savings are even greater. For instance, for a 100x100 pixel input field, a fully connected third-order network requires 1.6×10^{11} interconnections. If we represent this field as 10 fields of 10x10 coarse pixels, only 161,700 interconnections are necessary. The number of interconnections is decreased by a factor of ~100,000.

The relationship between number of coarse fields, n, input field size, IFS, and coarse field size, CFS, in each dimension is given by [7,9]

$$IFS = (CFS * n) - (n - 1). \tag{9}$$

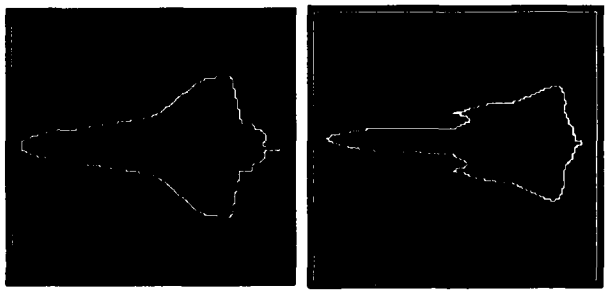


Figure 7: A binary edge-only representation of a Space Shuttle orbiter and an SR-71 aircraft, drawn in a 127x127 pixel window.

Training of the network proceeds in the usual way with one modification: the transfer function thresholds the value obtained from summing the weighted triangles over *all* coarse images associated with each training object. That is,

$$y = 1, \text{ if } \{\Sigma_n (\Sigma_j \Sigma_k \Sigma_l w_{jkl} x_j x_k x_l)\} > 0,$$
$$y = 0, \text{ otherwise,} \quad (10)$$

where j, k, and l range from one to the coarse pixel size squared, n ranges from one to the number of coarse fields, the x's represent coarse pixel values, and w_{jkl} represents the weight associated with the triplet of inputs (j, k, l). During testing, an input image is transformed into a set of coarse images. Each of these "coarser" vectors are then presented to the network and an output value determined using Eq. (10).

Software results: coarse-coded networks

We evaluated the coarse coding technique using an expanded version of the T/C problem. Implementing coarse coding, we increased the input image resolution for the T/C problem to 127x127 pixels using 9 fields of 15x15 coarse pixels. The network was trained on just two images: the largest "T" and "C" possible within the input field, and training took just five passes.

A complete test set of translated, scaled, and one degree rotated views of the two objects in a 127x127 pixel input field consists of ~135 million images. Assuming a test rate of 200 images per hour, it would take about 940 computer-months to test all possible views. Accordingly, we limited the testing to a representative subset consisting of four sets:

(1) All translated views, but with the same orientation and scale as the training images.
(2) All views rotated in-plane at 1° intervals, centered at the same position as the training images but only 60% of the size of the training images.
(3) All scaled views of the objects, in the same orientation and centered at the same position as the training images.
(4) A representative subset of approximately 100 simultaneously translated, rotated, and scaled views of the two objects.

The network achieved 100% accuracy on all test images in sets (1) and (2). Furthermore, the network recognized, with 100% accuracy, all scaled views, from test set (3), down to 38% of the original size. Objects smaller than 38% were all classified as C's. Finally, for test set (4), the network correctly recognized all images larger than 38% of the original size, regardless of the orientation or position of the test image.

A third-order network also learned to distinguish between practical images such as a Space Shuttle Orbiter versus an SR-71 aircraft (Figure 7) in up to a 127x127 pixel input field. In this case, training took just six passes through the training set, which consisted of just one (binary, edge-only) view of each aircraft. As for the T/C problem, the network achieved 100% recognition accuracy of translated and in-plane rotated views of the two images. Additionally, the network recognized images scaled to almost half the size of the training images, regardless of their position or orientation.

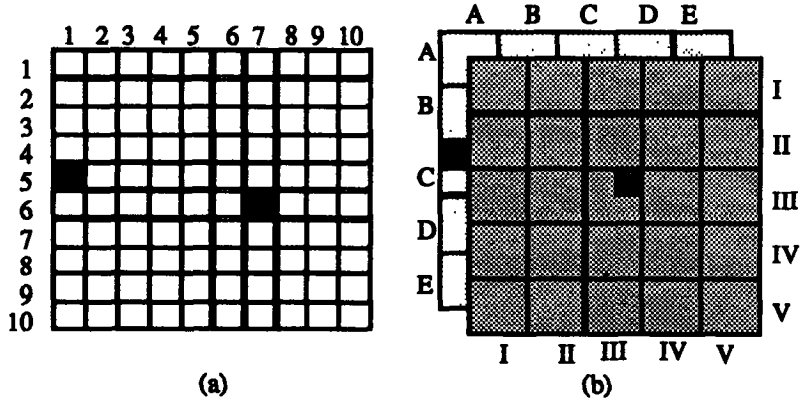


Figure 8: Example of a coarse-coded input field. (a) A 10x10 pixel input field. (b) Two fields of 5x5 coarse pixels.

The maximum input field resolution possible with coarse coded HONNs has not yet been reached. We ran simulations on the T/C problem coded with a variable number of 3x3 coarse pixels. A third-order network was able to learn to distinguish between the two characters in less than ten passes in an input field size of up to 4095x4095 pixels using 2047 fields. We expect a resolution of 4096x4096 is sufficient for most object recognition tasks. Notwithstanding, we also expect greater resolution is possible.

IV. TOLERANCE TO NOISE

All the demonstrations discussed so far showed the performance of HONNs in a noise-free environment. In this section, we discuss the recognition accuracy of HONNs with non-ideal test images. We consider white Gaussian noise and occlusion.

We evaluated the performance of HONNs with noisy images on two object recognition problems: an SR-71/U-2 discrimination problem and an SR-71/Space Shuttle discrimination problem. All simulations used a coarse-coded third-order network designed for a 127x127 pixel input field. We used 9 fields of 15x15 coarse pixels and a resolution of 10° for the angles α, β, and γ in Eq. (2), which allowed scale invariance over the range between 70% and 100% of the original image size. Each instantiation of the network was trained on just one binary, edge-only view of each object, as shown in Figure 10a, and training required less than ten passes through the training set.

The training sets were generated from 8-bit gray level images of actual models of the aircraft. The images were thresholded to produce binary images, and then edge detected using a digital Laplacian convolution filter with a positive derivative to produce the silhouettes shown in Figure 9a. For rotated and scaled views of the objects, the original gray level images were first scaled, then rotated, and then thresholded and edge-detected. Test images were positioned arbitrarily to validate the translation invariance of the network. Notice that the profiles of the SR-71 and Space Shuttle are somewhat similar whereas those of the SR-71 and U-2 are very different.

White Gaussian Noise

To test the tolerance of higher-order neural networks to white noise, each instantiation of the network (one for the SR-71/U-2 problem and one for the Shuttle/SR-71 problem) was tested on 1200 images generated by modifying the 8-bit gray level values of the original images using a Gaussian distribution of random numbers with a mean of 0 and a standard deviation of between 1 and 50. The noisy images were then geometrically distorted, binarized, and edge-enhanced. Typical test images which were correctly identified are shown in Figure 9b.

The results are summarized in Figure 10. The network performed with 100% accuracy for our test set for a standard deviation of up to 23 on the SR-71/U-2 problem and 26 on the Shuttle/SR-71 problem. For the similar images of the Shuttle and SR-71, the recognition accuracy quickly decreased to 75% at a σ of 30 and to 50% (which corresponds to no better than random guessing) for σ greater than 33. The SR-71/U-2 remained above 75% accuracy up to a σ of 35 (or ~14% of the gray level range) and gradually decreased to 50% at a σ of 40 (or ~16% of the gray level range). If we define "good performance" as greater than 75% accuracy, HONNs have good performance for σ up to 35 (or ~14% of the gray level range) for images with very distinct profiles and σ up to 30 (or ~12% of the gray level range) for images with similar profiles.

Occlusion

To test the tolerance of HONNs to occlusion, the two instantiations (one for the Shuttle/SR-71 problem and one for the SR-71/U-2 problem) of the third-order network built to be invariant to scale, in-plane rotation, and translation as described above were tested on occluded versions of the image pairs. We started with binary, edge-only images and added automatically-generated occlusions based on four variable parameters: the size of the occlusion, the number of occlusions, the type of occlusion, and the position of the occlusion. Objects used for occlusion were squares with a linear dimension between one pixel and twenty-nine pixels. The number of occlusion objects per image varied from one to four, and the randomly chosen type of occlusion determined whether the occlusion objects were added to or subtracted from the original image. Finally, the occlusions were randomly (uniform distribution) placed on the profile of the training images. The test set consisted of 10 samples for each combination of scale, rotation angle, occlusion size, and number of occlusions for a total of 13,920 test images per training image or 27,840 test images per recognition problem. Typical test images are shown in Figure 9c.

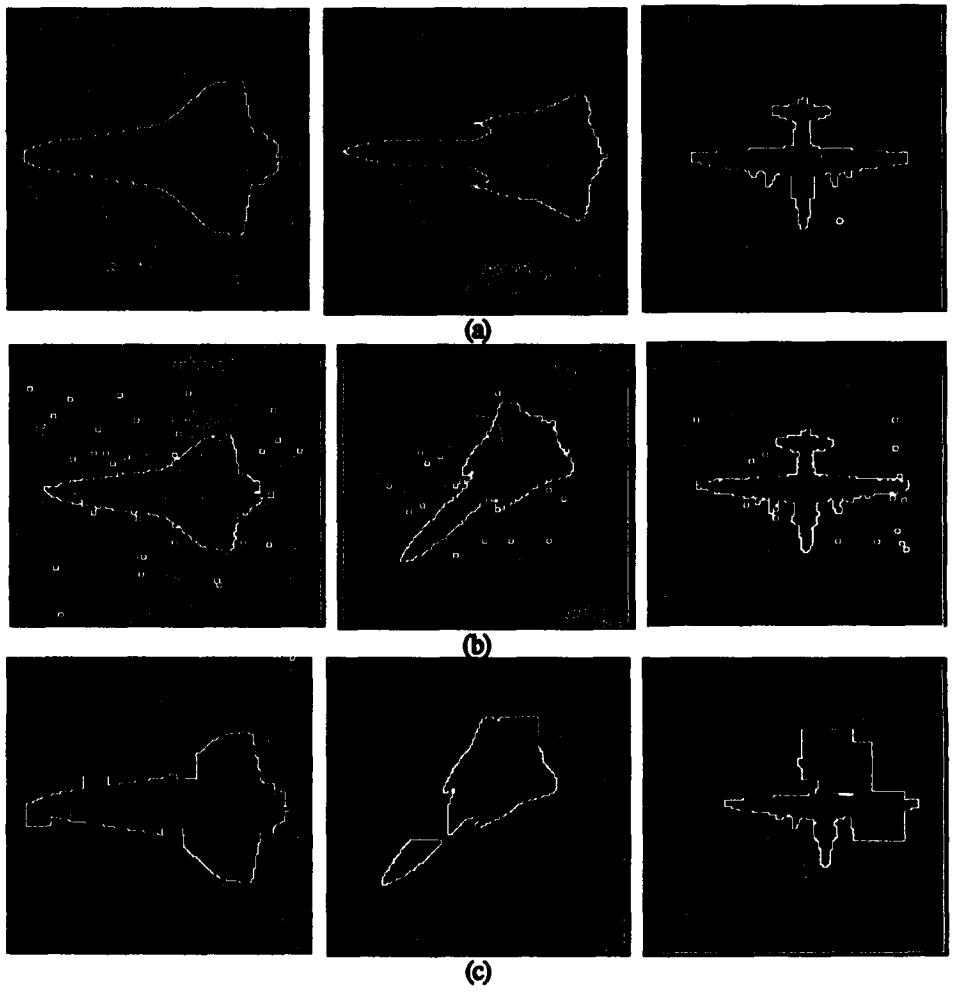


Figure 9: Training images in 127x127 pixel fields. (a) Binary edge-only training images of Space Shuttle Orbiter, SR-71, and U2. (b) Geometrically distorted and noisy test images correctly identified. (c) Geometrically distorted and occluded test images correctly identified.

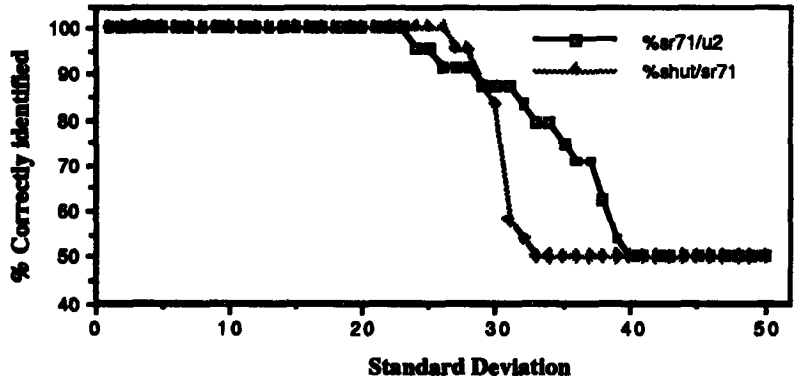


Figure 10: Tolerance of HONNs to white Gaussian noise. Each instantiation of a third-order network (one for the SR-71/U-2 problem and one for the Shuttle/SR-71 problem) was tested on 1200 test images generated by modifying the 8-bit gray level values of scaled, rotated and translated versions of the original training images using a Gaussian distribution of random numbers with a mean of 0 and a standard deviation between 1 and 50.

The performance of HONNs with occluded test images depends mostly on the number and size of occluding objects and to a lesser degree on the similarity of the training images. In the case of the Shuttle/SR-71 recognition problem, the network performed with 100% accuracy for our test set for one 16 pixel occlusion and up to four 10 pixel occlusions. It performed with better than 75% accuracy ("good performance") for up to four 19 pixel occlusions, three 21 pixel occlusions, two 24 pixel occlusions, and one 29 pixel occlusion.

For the SR-71/U-2 problem, the network exhibited good performance for the entire test set but achieved 100% accuracy only for one 4 pixel occlusion and up to four 3 pixel occlusions.

V. APPLICATION TO ROBOTIC VISION FOR MANUFACTURING

Vision processing is one of the most computationally intensive tasks required of an autonomous or semi-autonomous robot. A vision system based on a parallel implementation of a higher-order neural network can be used to perform one of the most difficult functions required of a general robotic vision system, distortion invariant object recognition, and can perform fast enough to keep pace with incoming sensor data. At Ames Research Center we have developed a robotic vision processing system to test concepts and algorithms for autonomous construction, inspection, and maintenance of space based habitats.

The benchmark task of the system is to allow a robot arm to identify and grasp an arbitrary tool moving in space with all six degrees of freedom without using any kind of cooperative marking techniques for the vision system. This is representative of one task required from the Flight Telerobotic Servicer (FTS) or the EVA Retriever, both of which are robots designed to operate in a weightless environment. A higher-order neural network can satisfy the first system task of object identification, after which other image processing sub-systems perform the tasks necessary to allow grappling.

We have tested a HONN-based vision system in the control of a Microbot robotic arm. The task was a subset of the benchmark task of allowing a robot arm to identify, track, and grasp an arbitrary tool without using any kind of cooperative marking techniques for the vision system. The robot arm carries a camera to observe the workspace below it, as shown in Figure 11.

The vision system task was to find one of a set of tools, as shown in Figure 12. The object set consists of five common tools and a structural component designed for automated in-space assembly. The work area is draped in black cloth to control the amount of background clutter. The robot was directed to look at each "bin" space in the work area, and to identify the tool located there. The tool could be located at any location within the bin, could be rotated in-plane. The camera height was not held constant, so the tools had varying apparent size. When the desired tool was found, a grappling operation was initiated.

This system also demonstrates the capability of HONN-based vision for a part/product identification task on a manufacturing assembly line. For example, parts on an assembly line passing below a camera could be quickly identified, regardless of their position, orientation, and (if need be) size.

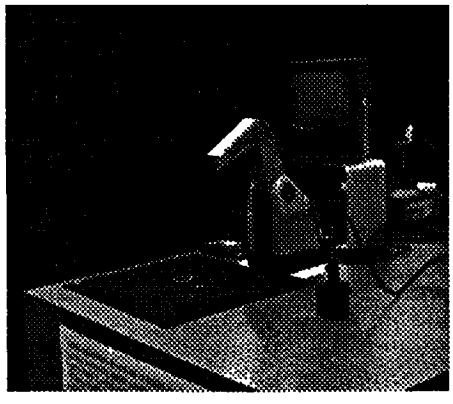


Figure 11: Photograph of the table top 5 degree-of-freedom Microbot arm and the work surface. The camera is attached to the writst of the arm

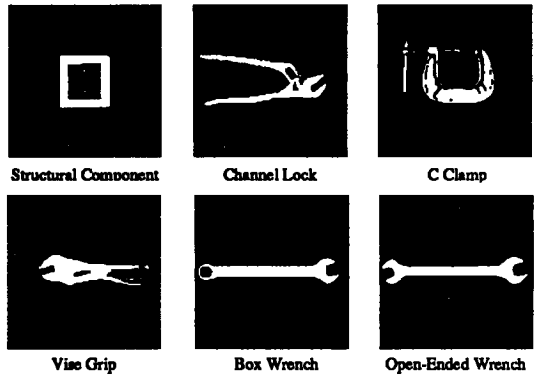


Figure 12: Binarized images of tools for recognition by the HONN vsion system. the images are edge-enhanced before being input to the HONN.

VI. CONCLUSIONS

We have shown that third-order neural networks can be trained to distinguish between two objects regardless of their position, angular orientation, or scale and achieve 100% accuracy on test images characterized by built-in distortions. Only one view of each object is required for learning and the network successfully learned to distinguish between all distorted views of the two objects in tens of passes, requiring only minutes on a Sun 3/60 workstation. In contrast, other neural network approaches require thousands of passes through a training set consisting of a much larger number of training images.

The major limitations of HONNs is that the size of the input field is limited because of the memory required for the large number of interconnections in a fully connected network. To circumvent this limitation, we developed a coarse coding algorithm which allows a third-order network to be used with a practical input field size of at least 4096x4096 pixels while retaining its ability to recognize images which have been scaled, translated, or rotated in-plane.

We explored the tolerance of higher-order neural networks (HONNs) to white Gaussian noise and to occlusion. We demonstrated that for images with an ideal separation of background/foreground gray levels, it takes a great amount of white noise in the gray level images to affect the binary, edge-only images used for training and testing the system to a sufficient degree that the performance of HONNs was seriously degraded. HONNs are also robust with respect to occlusion.

A third order neural network has been demonstrated in the laboratory for the control of a robot performing a typical manufacturing task of part identification. Our current research aims to extend the capabilities of this vision system by training a third order to recognize out-of-plane rotated versions of a training object. With scale, position, and in-plane rotation invariance built into the architecture, and out-of-plane invariance learned, a full six degree of freedom vision system can be achieved. In addition, we are working on a implementation of a third order network on a parallel processor, which will allow the identification of objects in a 128x128 pixel image at full video (60 Hz) rates.

All our current software runs on any Sun workstation, either a Sun 3/60 or a SPARC system. This software will soon by available through COSMIC, the U.S. Government's software distribution facility.

VII. REFERENCES

[1] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[2] Chen, C.H., Statistical Pattern Recognition. Washington: Hayden, 1973.

[3] G.L. Giles and T. Maxwell, "Learning, invariances, and generalization in high-order neural networks," Applied Optics, 26, 1987, pp. 4972-4978.

[4] G.L. Giles, R.D. Griffin, and T. Maxwell, "Encoding geometric invariances in higher-order neural networks," Neural Information Processing Systems, American Institute of Physics Conference Proceedings, 1988, pp. 301-309.

[5] M.B. Reid, L. Spirkovska, and E. Ochoa, "Simultaneous position, scale, and rotation invariant pattern classification using third-order neural networks," International Journal of Neural Networks, 1, 1989, pp. 154-159.

[6] M.B. Reid, L. Spirkovska, and E. Ochoa, "Rapid Training of Higher-Order Neural Networks for Invariant Pattern Recognition," Proceedings of Joint International Conference on Neural Networks, Washington, D.C., June 18-22, Vol. 1, 1989, pp. 689-692.

[7] L. Spirkovska and M.B. Reid, "Application of Higher-Order Neural Networks in the Position, Scale, and Rotation Invariant Object Recognition Domain," invited submission to Sixth Generation Computing, Wiley, New York, to be published, 1992.

[8] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, Parallel Distributed Processing, Vol. 1, Ch. 8, pp. 348-352. Cambridge, Massachusetts: MIT Press, 1986.

[9] L. Spirkovska and M.B. Reid, "Coarse-Coded Higher-Order Neural Networks for PSRI Object Recognition," submitted to IEEE Transactions on Neural Networks.

N92-22714

CONSTRAINT-BASED SCHEDULING

Monte Zweben
Artificial Intelligence Research Branch
NASA Ames Research Center
Moffett Field, CA 94035
Mail Stop 244-17
zweben@ptolemy.arc.nasa.gov

ABSTRACT

The GERRY scheduling system developed by NASA Ames with assistance from the Lockheed Space Operations Company, and the Lockheed Artificial Intelligence Center, uses a method called constraint-based iterative repair. Using this technique, one encodes both hard rules and preference criteria into data structures called constraints. GERRY repeatedly attempts to improve schedules by seeking repairs for violated constraints. The system provides a general scheduling framework which is being tested on two NASA applications. The larger of the two is the Space Shuttle Ground Processing problem which entails the scheduling of all the inspection, repair, and maintenance tasks required to prepare the orbiter for flight. The other application involves power allocation for the NASA Ames wind tunnels. Here the system will be used to schedule wind tunnel tests with the goal of minimizing power costs. In this paper, we describe the GERRY system and its application to the Space Shuttle problem. We also speculate as to how the system would be used for manufacturing, transportation, and military problems.

INTRODUCTION

Efficient scheduling is crucial for manufacturing companies that must balance limited production resources against challenging order requests. Airlines and package delivery companies must schedule large fleets of vehicles coordinating transportation goals with maintenance goals but must also be adaptable to external forces such as weather and equipment failure. The DoD also faces daunting scheduling problems ranging from logistics transport problems to mission planning problems.

NASA also faces complex scheduling problems including telescope observation scheduling, spacecraft crew scheduling, and spacecraft mission planning. Our research is motivated by the Space Shuttle Ground Processing problem. Ground processing entails the inspection, testing, and repair activities required to prepare a Space Shuttle for launch at the Kennedy Space Center (KSC) in Florida.

This paper describes a scheduling algorithm that is being used to schedule shuttle ground processing but is also applicable to the other scheduling problems alluded to above. First we present our definition of a scheduling problem and then describe our scheduling method. After presenting the general approach we describe how it is used to solve the Space Shuttle problem and then briefly describe how it can be adapted to other real-world problems.

SCHEDULING

In this section we define the scheduling problem beginning with a simplified version and evolving to a more realistic definition.

General Problem

Generally scheduling systems are provided a set of activities, relationships between these activities (such as predecessor-successor requirements), resource requirements for each task (i.e., how much of what kind of resources are necessary), and a set of deadlines or milestones. With this input, scheduling systems determine start and end times as well as an assignment of resources to each activity such that: 1) the relationships between tasks are preserved, 2) no resource is over-allocated (i.e., at no time does the demand for a

resource exceed its supply), and 3) all milestones are met.

For example, consider a Space Shuttle repair scenario where each Space Shuttle Main Engine needs to be inspected, removed, repaired, re-installed, and tested, in that order. The tasks associated with different engines are unrelated meaning that any task in support of one engine could simultaneously occur with the tasks in support of a different engine. Assume that each task requires 10 technicians, an engineer, and a safety inspector. Suppose there were only 15 technicians on call for each shift. In this case, no two activities would be able to occur in parallel because together they require 20 technicians and there were only 15 available. If there were more technicians the system would place tasks in parallel in order to meet the milestone.

Consequently, a scheduler would determine activity start times that sequence the activities completely serially because any two activities' demand exceeds the supply of technicians.
In summary, scheduling systems search through the space of possible start times and resource assignments with the goal of finding an assignment that satisfies all domain constraints. These constraints include milestones, resource capacities, and temporal relationships.

Optimizing and Satisficing

Most scheduling systems simply find an acceptable schedule and then terminate. They are not necessarily concerned with finding the best schedule that satisfies the constraints. In many domains, there is great variability in the quality of schedules that satisfy constraints. For example, an organization might want to find the schedule that uses the minimal amount of overtime labor, or one that minimizes the overall flowtime of a schedule. Unfortunately, deriving the *optimal* schedule is a time consuming process that requires a great deal of combinatoric search. In most cases, near-optimal solutions are sufficient. The process of problem-solving with the goal of finding near-optimal solutions is called *satisficing*"[Simon]. The satisficing algorithm presented in the next section continues to search after finding a schedule that merely satisfies constraints, in order to find better quality schedules according to stated optimization criteria.

State Conditions

Most scheduling systems reason about the changing availability of resources over time but few track the changes of arbitrary conditions. State conditions are attributes of the scheduling problem that change with time. The tasks of a scheduling problem are constrained by these conditions and occasionally the activities change the values of the conditions. Examples include the position of switches and other mechanical parts, the readings of sensors, and the location of objects. Schedulers that handle state conditions must provide a language to specify the additional *state constraints* and to specify the *effects* that tasks have on state conditions.

Examples of state conditions in the Space Shuttle scheduling problem include the position of the payload bay doors, the status of the orbiter's hydraulics system, and whether an area adjacent to the orbiter is hazardous. Examples of state constraints include the rule that no task may take place in a hazardous area. Additionally, some activities require orbiter hydraulics while others require the hydraulics to be off. Likewise, certain activities require the payload bay doors to be in one of their three main positions. Activities also change these conditions. Some activities result in opening or closing doors and turning hydraulics on or off. Similarly, hazardous operations cause the areas surrounding their respective work areas to be considered hazardous thus delaying any other operations that must share those areas.

Our system supports the modeling of state conditions and provides a language for state constraints and task state effects however, details of this language are beyond the scope of this paper. It suffices to say that the satisficing search mechanism presented below considers state constraints and state effects as it schedules.

Pre-emptive Scheduling

Pre-emption is the process of temporarily suspending activities and resuming them later. Pre-emption can be caused for a number of reasons. In a telescope observation scheduler, the system might interrupt an activity when a more important and rare astronomical event arises. Activities could also be suspended to allow more contentious activities to execute in their limited windows of opportunity. These are examples of *flexible pre-emption*. The Space Shuttle problem requires a more restricted type of pre-emption called *fixed pre-emption*.

Fixed pre-emption is the suspension and resumption of activities according to a strict calendar. In the Shuttle domain the calendar corresponds to work shifts. Some activities can be worked all shifts, every day, while others have certain restrictions such as no weekends, only third shift, or only first shift.

To handle this sort of pre-emption our system requires a calendar for each task that indicates how it is to be pre-empted or split into smaller pieces. For example, suppose a task that requires 12 hours work is assigned a first shift, no weekends calendar. If the task begins at 8:00 A.M. Monday, it will be suspended at 4:00 P.M. that day, then resumed Tuesday morning at 8:00 A.M., and then finally completed at noon. Thus the task spans two calendar days. Suppose however that the task began Friday. It will then terminate Monday thus spanning four calendar days.

Pre-emption greatly complicates scheduling because of the way it interferes with resources and state conditions. Whenever a new time is considered for a task, the task must be split according to the calendar. However, it is sometimes inappropriate for the state and resource constraints to be valid for the entire period of the pre-empted tasks. For example, the resource needs that correspond to human labor should not be required during the suspended periods of the task's duration. In other words, it makes no sense for employees to be standing around idle. In these cases, the resource and state constraints must be *inherited* to the split tasks thus avoiding the idle periods. Other constraints can remain active throughout the duration of the task. An example of these include a resource request for a heavy piece of equipment that requires significant assembly. The equipment usually remains in the work area, unavailable to others because of the overhead required to set it up.

Our system allows the user to designate which resource requests and which state constraints and effects are to remain valid throughout the suspended period and which ones are valid only during active periods.

CONSTRAINT-BASED SCHEDULE REPAIR

In this section, we present the search method used by the GERRY scheduling system. The system allows the user to specify a set of *tasks*, a set of *state conditions*, and a set of *resources*.

Tasks have start and end times, resource requests, resource assignments, work durations, and calendars.

Resource pools are defined by the user and have a corresponding maximum capacity. For example, in the Space Shuttle domain there might be a pool of 20 technicians or three pools of 5 forklifts.

State conditions are also provided by the user along with the initial values for each condition. For example, the right-hand payload bay door with an initial value of *closed*.

Input

- Task Data - For each task, the following information is provided:
  - work duration - amount of active work time required for the task to complete.
  - calendar - the pre-emption times for the task.
  - resource requests - the list of resource types and quantities necessary.

- Resource Data - For each resource pool, the following information is provided:
  - type - the name of the resource category that the pool is classified as.
  - capacity - the maximum amount of the pool that can be simultaneously assigned.

- State Condition Data - For each state condition, the following information is provided:
  -initial value - the value for a condition which persists until a task changes it.

Output

For each task, the following information is determined:
- start time - the beginning of the task.
- end time - the finish of the task.
- resource assignments - the actual resources chosen.

The rules and preferences that schedules must observe are captured by *constraints*. Constraints are relationships that are desired by the user and are composed of the following items:

- Arguments - the tasks, resources, or state conditions that are related to each other.
  Example: a task and a resource pool are arguments to a resource capacity constraint.

- Penalty - a score of how poor the arguments are with respect to the constraint.
  Example: if the resource pool argument were overallocated during the time of the task, then the penalty of the resource capacity constraint would be high.

- Weight - a number reflecting the importance of the constraint.
  Example: resource capacity constraints for scarce resources such as expensive equipment would have higher weights.

- Repairs - suggested schedule modifications that are intended to improve the penalty.
  Example: move tasks that are involved in an overallocation to a time where more of the resource is available.

Loosely speaking, the penalty is analogous to the amount of money one would have to pay with respect to the current assignment of times and resources. The weight of the constraint reflects its importance when compared to other constraints. Repairs are methods for changing the schedule, either by substituting resources, or by moving, adding, or deleting activities.

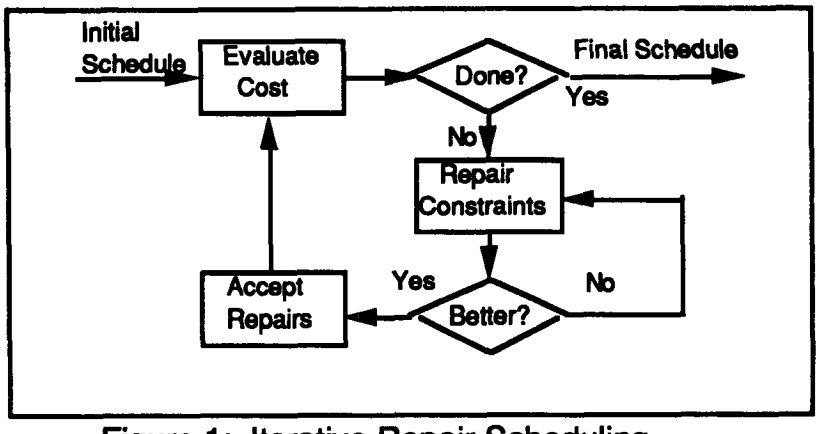


Figure 1: Iterative Repair Scheduling

Figure 1. presents our simple iterative repair algorithm. See [Zweben et. al.][Zweben1 et. al.] for details. The system begins with an initial schedule and then initiates a repair loop.[1] If the problem posed to the system is a rescheduling problem then the initial schedule is the schedule with changes imposed by the user. If the system is scheduling from scratch, then all tasks are placed at their earliest possible start times while preserving temporal constraints. This is accomplished with a well known polynomial (i.e., efficient) algorithm [Davis, Waltz].

In the repair loop, the system calculates the cost of the solution. This calculation is simply the sum of each constraint's penalty multiplied by its weight. If the cost is below a threshold set by the user, the search terminates.[2] Otherwise, a cross-section of the highly penalized constraints are repaired. We often refer to these constraints as the violated constraints because their penalties exceed a certain threshold.

In short, the system simply starts with a schedule and isolates the violated constraints. Then it moves tasks around and substitutes resources as suggested by the repairs embodied in the violated constraints. It accepts the new schedule if the new cost is lower than the previous cost. If the repaired schedule is worse than previous one, it is rejected and new repairs are attempted on the previous schedule.[3] The system continues this process until the cost of the solution is acceptable to the user, or the system is terminated by the user. The system also terminates if a certain number of iterations have been tried.

SYSTEM FUNCTIONALITY

User Interface Overview

GERRY allows both manual and automatic scheduling thus requiring a sophisticated user interface. The user instructs the interface to display a chart. The chart library is extensible and allows the user to define different views of the schedule. For example, the user could ask for a time-line (i.e., Gantt Chart) and a resource profile (i.e., a histogram of resource usage over time) to be displayed for every task. Alternatively, the user could require a time-line and state condition profile (i.e., a histogram of state conditions over time) for a specific set of tasks. Figure 2. is part of a Space Shuttle schedule with a time-line and resource profile. Figure 3. shows the same schedule (with hourly units) at a coarser level of resolution. Zooming in and out of different levels of resolution is accomplished by clicking in the upper right hand boxes of the chart. This convention was adopted from the COMPASS scheduling system of Barry Fox at McDonnell Douglass in Houston, Texas.

---

[1] Similar repair techniques are used in OMP [Biefeld, et. al.] and in the work on the MIN-CONFLICTS heuristic [Minton, et.al.].

[2] The search also terminates if the system exceeds the iteration bound imposed by the user.

[3] Actually the system sometimes accepts worse schedules in order to break out of situations called local minima.. In a local minimum, any repair leads to a worse schedule, but subsequent repairs could improve the schedule so that it is eventually superior. This technique is called simulated-annealing and was originally reported in [Kirkpatrick].

When the chart is displayed each of the activities and histograms are mouse-sensitive. One can drag a task to a new location, modify its status as pending, active, or complete, and ask for a the list of resources that the task uses. The status of an activity is reflected by the shading of the displayed bar. If the activity is shaded black, it is complete. Ongoing activities are outlined in black (but not shaded). Unshaded activities that do not have an outline are pending.

In addition to status, shading is also used to reflect the danger level of the task. If an activity is shaded with a cross-hatch pattern then it is considered hazardous.

The interface supports many task look-up methods. For example, one can scroll to a point in a chart where a particular task begins, one can scroll to an over-allocation, or one could simply use scroll bars.

Also included in the interface are a form editor and a temporal constraint grapher. The form editor, shown in Figure 4., is simply a mechanism to enter new activities and constraints. The grapher, shown in Figure 5., allows the user to inspect the complex temporal relationships between tasks. The grapher works in a demand-driven manner instead of cluttering the display with the entire schedule's graph. One clicks on a task and the graph expands from that point on.

Schedule Monitoring and Rescheduling

While GERRY can be used as a planning tool for future schedules, its strength is in its ability to monitor schedule execution and adapt to the schedule changes imposed by elements outside the system's control. Users modify tasks by changing their status, dragging them around, changing their constraints, and by adjusting task durations. Users can also add and delete tasks.

One of the most important functions of the user interface is the ability to alert the user to the ramifications of their changes. There are three main charts used to inform the user of what they have done: 1) a before-and-after chart, 2) a constraint violation summary chart, and 3) a constraint violation problem report. The before-and-after chart, shown in Figure 6., reports all the tasks that have been changed by indicating their new position and their old position. The constraint violation summary chart, shown in Figure 7., is a list of the current constraint violations. By clicking on one of the violations in the constraint violation summary chart, a constraint violation problem report appears that explains the conflict. For example, the constraint violation problem report shown in Figure 8. explains why a particular resource capacity constraint is violated. Only the tasks that request the resource during the interval of the violation are displayed and the interval is shaded in color.

After changes are given to the system, the user can manually resolve the outstanding violations or ask the system to use the iterative repair method. When the automated method terminates (or is interrupted) it reports a before-and-after chart. If there are outstanding violations, then a constraint violation summary chart is also displayed.

PROBLEM DOMAINS

Space Shuttle Ground Processing

In the Space Shuttle Ground Processing domain we start with schedules provided by an existing project management tool used at the Kennedy Space Center. It uses the critical path method (CPM) to schedule activities at the earliest possible times. These schedules are used by Flow Managers who have the responsibility to prepare the orbiter in time for the designated launch date. The data sets start with about 300-400 activities that expand into thousands of split tasks and constraints. Our project team attends KSC schedule meetings and updates the schedule accordingly. Currently, we are in the process of delivering new schedules to the flow managers and beginning to use the constraint-based repair method to optimize the schedule. Below we enumerate the constraints used for this application.

The constraints for this application include:

1. Resource Capacity Constraints

- Arguments: - Start of a task
  - End of a task
  - Resource Pool

- Penalty: - Constraint is violated when the pool is over-allocated during the task. Example: 3 tasks in parallel all need a technician but only 2 are on call.

- Repair: - Strategy 1: Substitute
  Assign a resource pool of the same type that is not over-allocated.

  - Strategy 2: Move
  Move one of the tasks contending for the resource to the next time when there is a sufficient amount of the resource.

  To decide which task to move the following heuristics are used:

  -Heuristic 1: Fitness
  Prefer to move tasks that use an amount of the resource that is close to the amount over-allocated.
  -Heuristic 2: Slack
  Avoid moving tasks that have little slack between their earliest and latest start times.[4]
  -Heuristic 3: Dependents
  Avoid moving tasks with temporal dependents (e.g., postrequisites).
  -Heuristic 4: Priority
  Avoid moving high priority tasks.
  -Heuristic 5: In Process
  Avoid moving tasks that have begun.
  -Heuristic 6: Proximity
  Avoid moving tasks that are to begin soon.

2. State Constraints

Arguments: - Start of a task
- End of a task
- State Condition
- Required State

Penalty: - Constraint is violated when the condition does not reflect the required state during the task.
Example: The payload bay doors are closed for a task that requires them to be 160 degrees open.

---

[4] Slack time indicates the amount of time a task could slip before it affects the milestone. This measure is calculated from the CPM algorithm mentioned earlier.

Repair: - Strategy 1: Move
Move the task to the next time where the state condition reflects the desired state.

3. Milestone Constraints

Arguments: - End of a task
- Due Date

Penalty: - Constraint is violated when the end of the task is completed later than the given date.

Repair: - Strategy 1: Move
Move the task back earlier, before the given time.

Currently we lack the domain knowledge that would distinguish between the importance of these constraints so they all have the same weight. The system uses these constraints (and their corresponding repairs) to minimize missed launch dates (via milestone constraints) and to minimize over-allocation of KSC personnel (via resource constraints) while maintaining the correct orbiter configurations (via state constraints).

In the near future we intend to include another constraint that demonstrates the flexibility of our system. This new constraint will inform the system to minimize labor costs by avoiding overtime labor on the weekend.

4. Weekend Constraints

Arguments: - Global constraint.

Penalty: - Constraint is violated when a large number of tasks intersect the weekend.

Repair: - Strategy 1: Move
Move the tasks with sufficient slack time off the weekend.

Manufacturing Problems

In job-shops, there are resources such as machines and human operators. Similar to pre-determined launch dates in the NASA domain, job shops have order due dates. In job shops, each machine has to be set up correctly depending upon the task at hand. Typically jobs follow a process plan that is fairly well known in advance. There are very similar optimization criteria in this domain as there are in the Space Shuttle domain. In fact, the constraints described above are usually applicable. Additional constraints would also be written that would modify the schedule to minimize the number of machine set-ups required thus minimizing flow time. Additionally, constraints that minimize the amount of work-in-process inventory would be incorporated. We claim that a knowledge engineer could easily do this without writing another program but rather simply writing new constraints.

Airline, Trucking, and Parcel Service Problems

In the transportation sector, large fleets of vehicles must be scheduled on a daily basis. These operations are stricken with unexpected events such as unpredicted malfunctions and malevolent weather. When these events occur, it is crucial to get back on track minimizing impact to the original schedule.

In transportation problems there are additional decision variables that constrain the schedule which include the start and end locations of any task and the speed that one will travel between those locations. Constraints would be added that relate the locations, speed, and duration of the task. Additionally the quantity of certain resource requests must be constrained by the duration. For example, the amount of fuel required by an aircraft is dependent upon how long the plane will travel. Constraints that serve to minimize fuel and delays, while observing safety constraints would be added to the constraints discussed above.

Military Problems

Many military problems resemble transportation problems but with targeting and probability of success factors added. The tasks are generally trips from one's bases to the enemy's targets (and hopefully back home again). In addition to the transportation constraints discussed above, constraints that model the appropriateness of various aircraft and ordnances for targets would be required.

Power Utilization

Ames Research Center is also deploying GERRY to minimize power consumption of the Ames wind tunnels. The rates that local utilities charge NASA are based upon the season, time of day, and quantity of power used. Therefore the wind tunnel test schedule can greatly affect energy costs. Ames will use GERRY to adjust the wind tunnel test schedule to minimize its power costs but maintaining the deadlines imposed by those who need the tunnels. Constraints are used to penalize schedules of high cost while repairs move tasks to lower these costs.

SUMMARY

We have developed a framework for scheduling called constraint-based iterative repair. This framework supports complex scheduling problems where satisficing is required. GERRY, the system based upon this framework, is operational and is being deployed at the Kennedy Space Center in Florida in support of Space Shuttle Ground Processing. The system uses the optimization criteria encoded as constraints to find near-optimal schedules. We claim that our approach is amenable to other problems faced within industry and government and welcome others to apply it.

Acknowledgements

Thanks to the entire GERRY project team at NASA Ames, the Lockheed Space Operations Company, and the Lockheed Artificial Intelligence Center. Also thanks to Martin Cohen for his careful review of this paper.

BIBLIOGRAPHY

[Biefeld, et. al.] Biefeld, E., and Cooper, L., Bottleneck Identification Using Process Chronologies, In *Proceedings of IJCAI-91*, 1991.

[Davis] Davis, E., Constraint Propagation with Interval Labels, *Artificial Intelligence*, 32(4), 1987.

[Kirkpatrick] Kirkpatrick, S., Gelatt Jr., C., Vecchi, M., Optimization by Simulated Annealing, *Science*, 220, 1983

[Minton, et. al.] Minton, S., Philips, A., Johnston, M., Laird, P., Solving Large Scale CSP and Scheduling Problems with a Heuristic Repair Method, In *Proceedings of AAAI-90.*

[Simon] Simon, H., *The Sciences of the Artificial*, MIT Press, 1969

[Waltz] Waltz, D. Understanding Line Drawings of Scenes with Shadows, In P. Winston, editor, *The Psychology of Computer Vision*, McGraw-Hill 1975.

[Zweben, et.al.] Zweben, M., Deale, M., and Gargan, B., Anytime Rescheduling, In *Proceedings of the DARPA Workshop on Innovative Approaches to Planning and Scheduling*, 1990.

[Zweben1 et.al.] Zweben, M., Deale, M., and Gargan, B., An Empirical Study of Rescheduling Using Constraint-Based Simulated Annealing, In *Proceedings of the IJCAI-91 Workshop on Production Planning and Scheduling*, 1991.

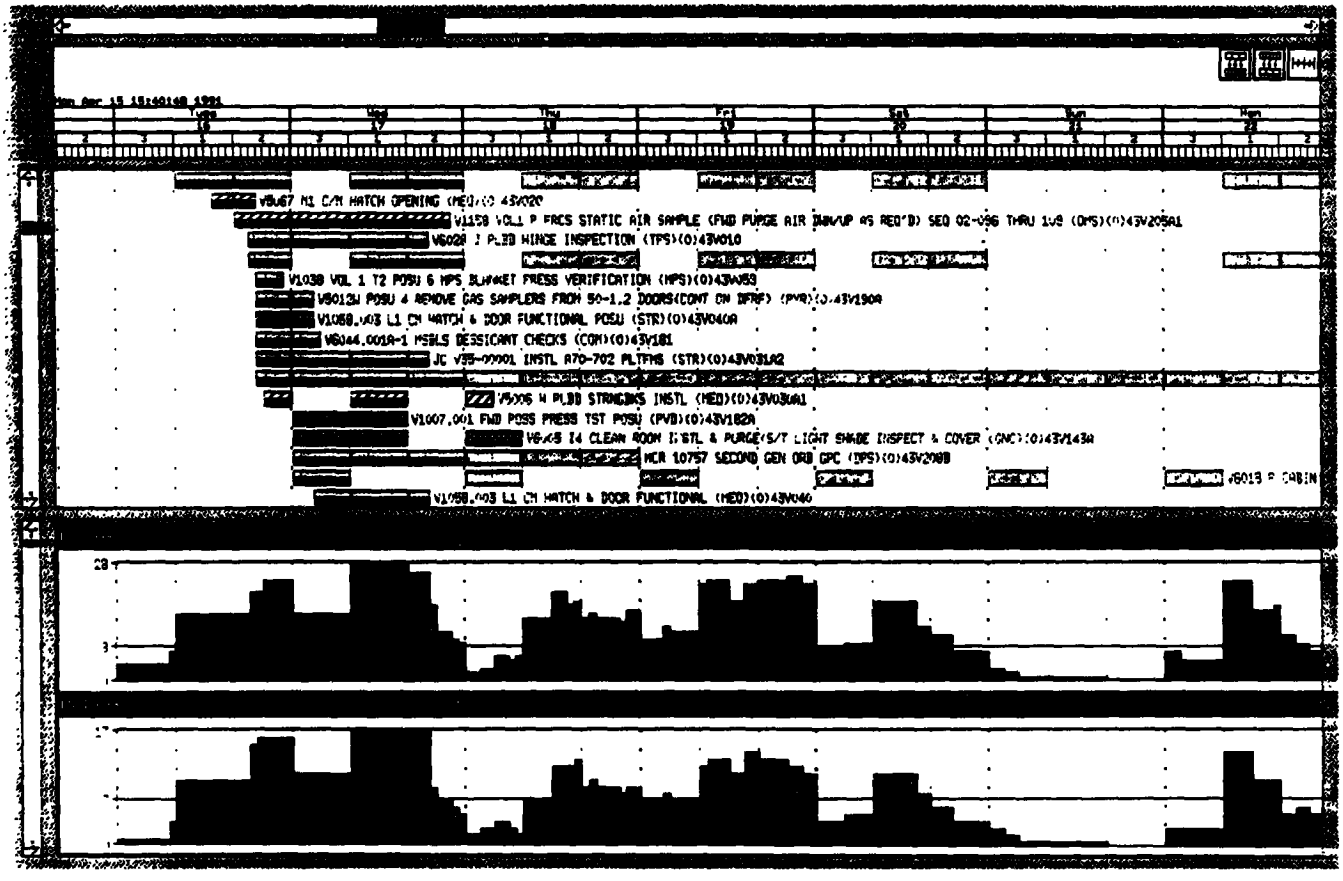


Figure 2.

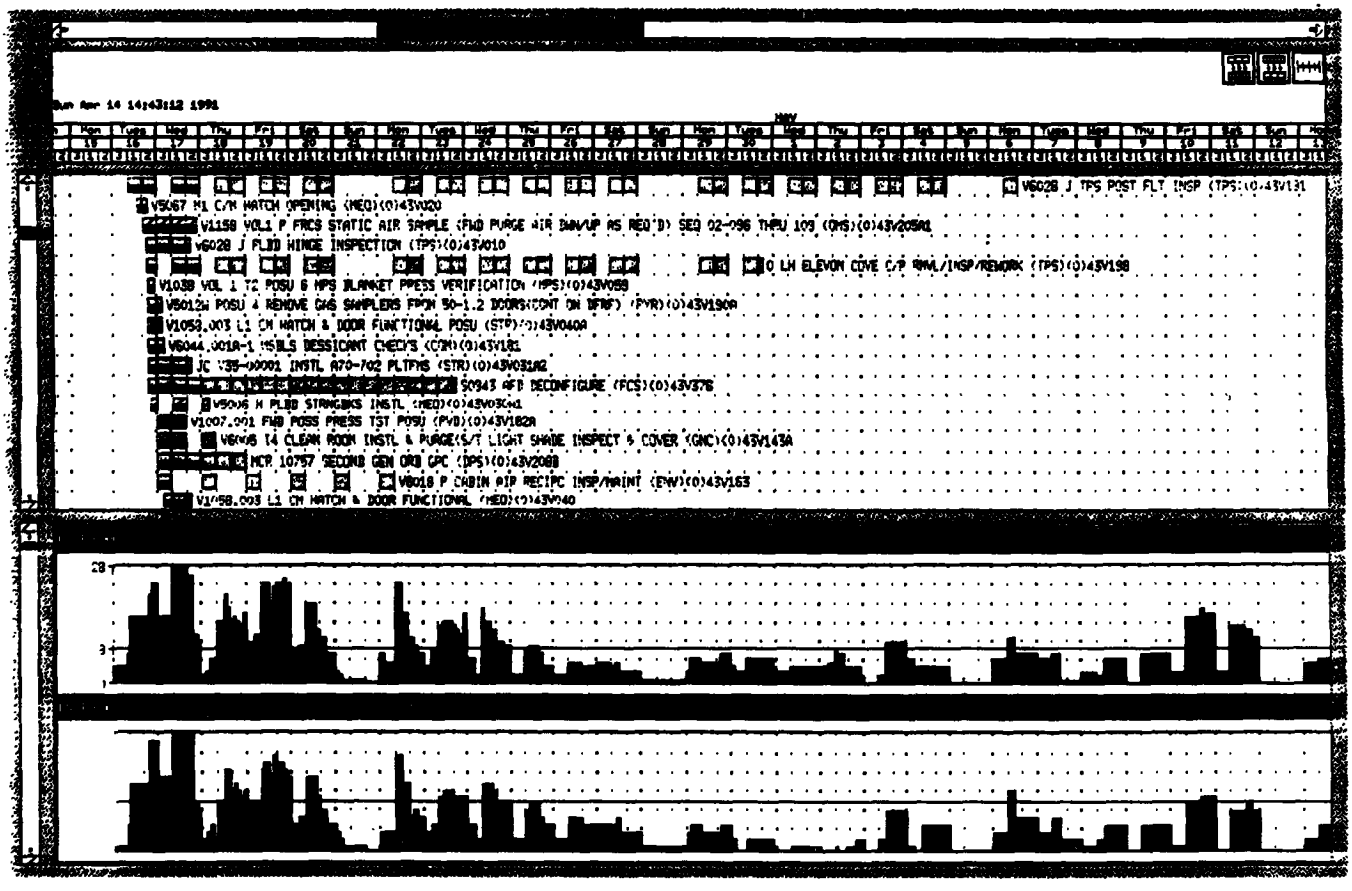


Figure 3.

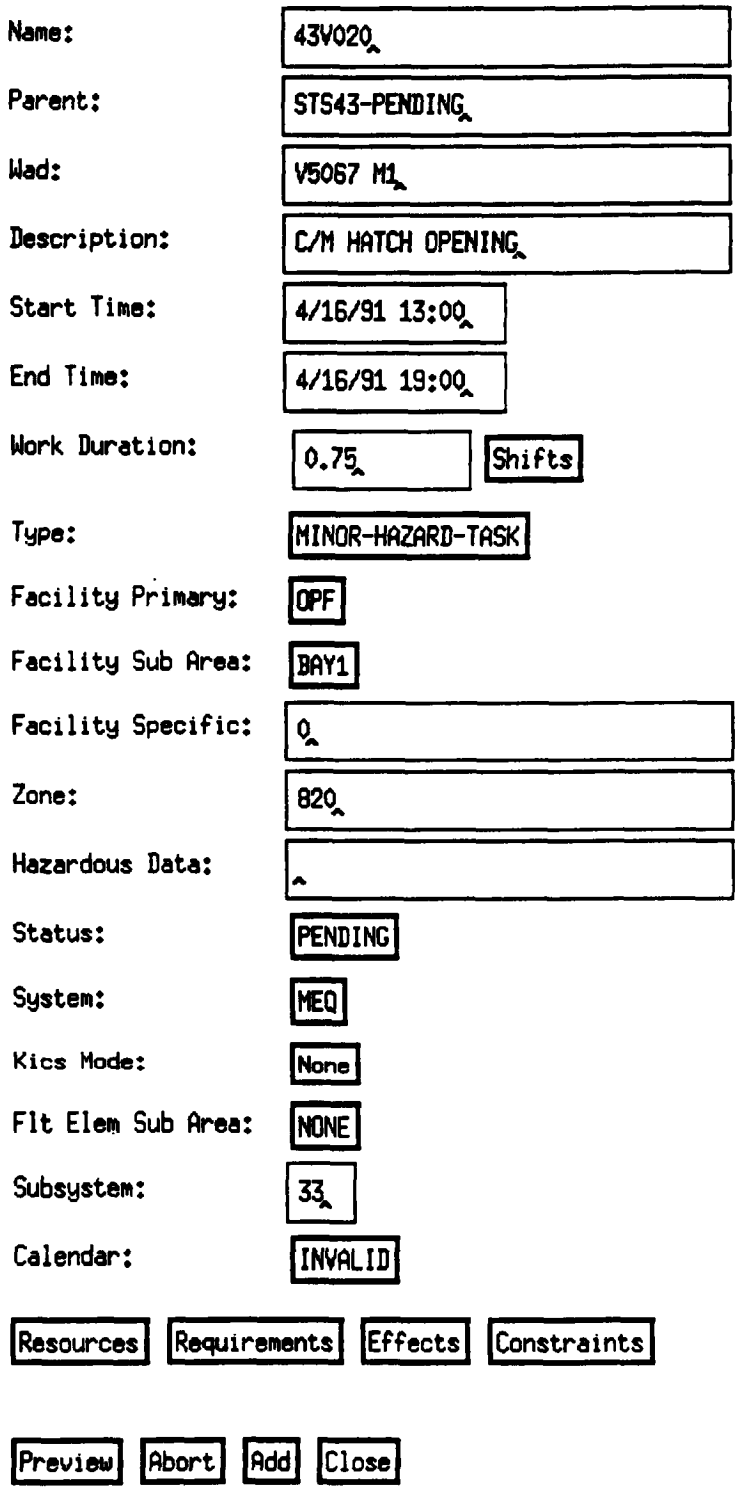

Name:
43V020
Parent:
STS43-PENDING
Wad:
V5067 M1
Description:
C/M HATCH OPENING
Start Time:
4/16/91 13:00
End Time:
4/16/91 19:00
Work Duration:
0.75
Shifts
Type:
MINOR-HAZARD-TASK
Facility Primary:
OPF
Facility Sub Area:
BAY1
Facility Specific:
0
Zone:
820
Hazardous Data:
Status:
PENDING
System:
MEQ
Kics Mode:
None
Flt Elem Sub Area:
NONE
Subsystem:
33
Calendar:
INVALID
Resources
Requirements
Effects
Constraints
Preview
Abort
Add
Close
Errors:

Figure 4.

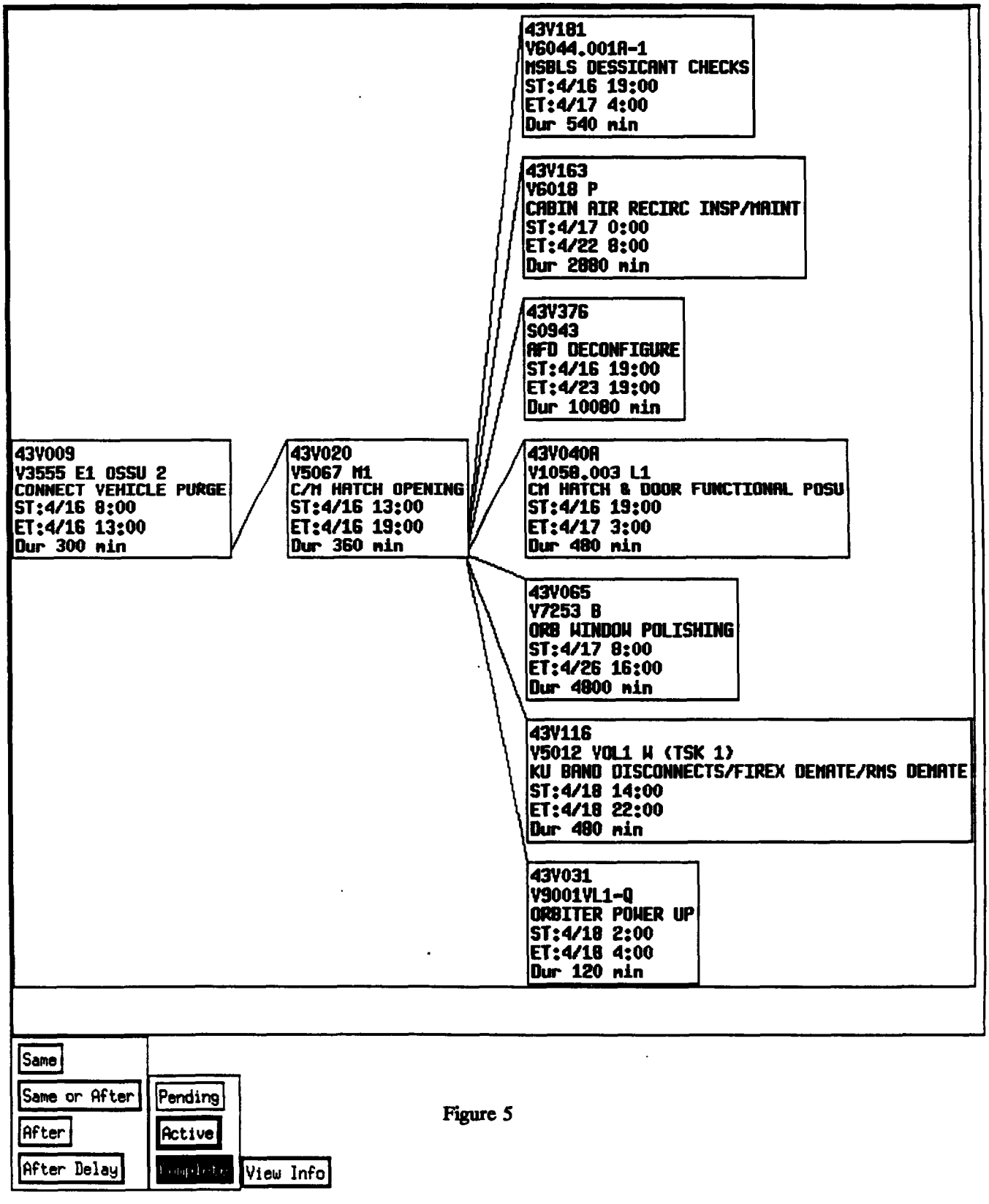

43V181
V6044.001A-1
MSBLS DESSICANT CHECKS
ST:4/16 19:00
ET:4/17 4:00
Dur 540 min
43V163
V6018 P
CABIN AIR RECIRC INSP/MAINT
ST:4/17 0:00
ET:4/22 8:00
Dur 2880 min
43V376
S0943
AFD DECONFIGURE
ST:4/16 19:00
ET:4/23 19:00
Dur 10080 min
43V009
V3555 E1 OSSU 2
CONNECT VEHICLE PURGE
ST:4/16 8:00
ET:4/16 13:00
Dur 300 min
43V020
V5067 M1
C/M HATCH OPENING
ST:4/16 13:00
ET:4/16 19:00
Dur 360 min
43V040A
V1058.003 L1
CM HATCH & DOOR FUNCTIONAL POSU
ST:4/16 19:00
ET:4/17 3:00
Dur 480 min
43V065
V7253 B
ORB WINDOW POLISHING
ST:4/17 8:00
ET:4/26 16:00
Dur 4800 min
43V116
V5012 VOL1 W (TSK 1)
KU BAND DISCONNECTS/FIREX DEMATE/RMS DEMATE
ST:4/18 14:00
ET:4/18 22:00
Dur 480 min
43V031
V9001VL1-Q
ORBITER POWER UP
ST:4/18 2:00
ET:4/18 4:00
Dur 120 min
Same
Same or After
After
After Delay
Pending
Active
View Info

Figure 5

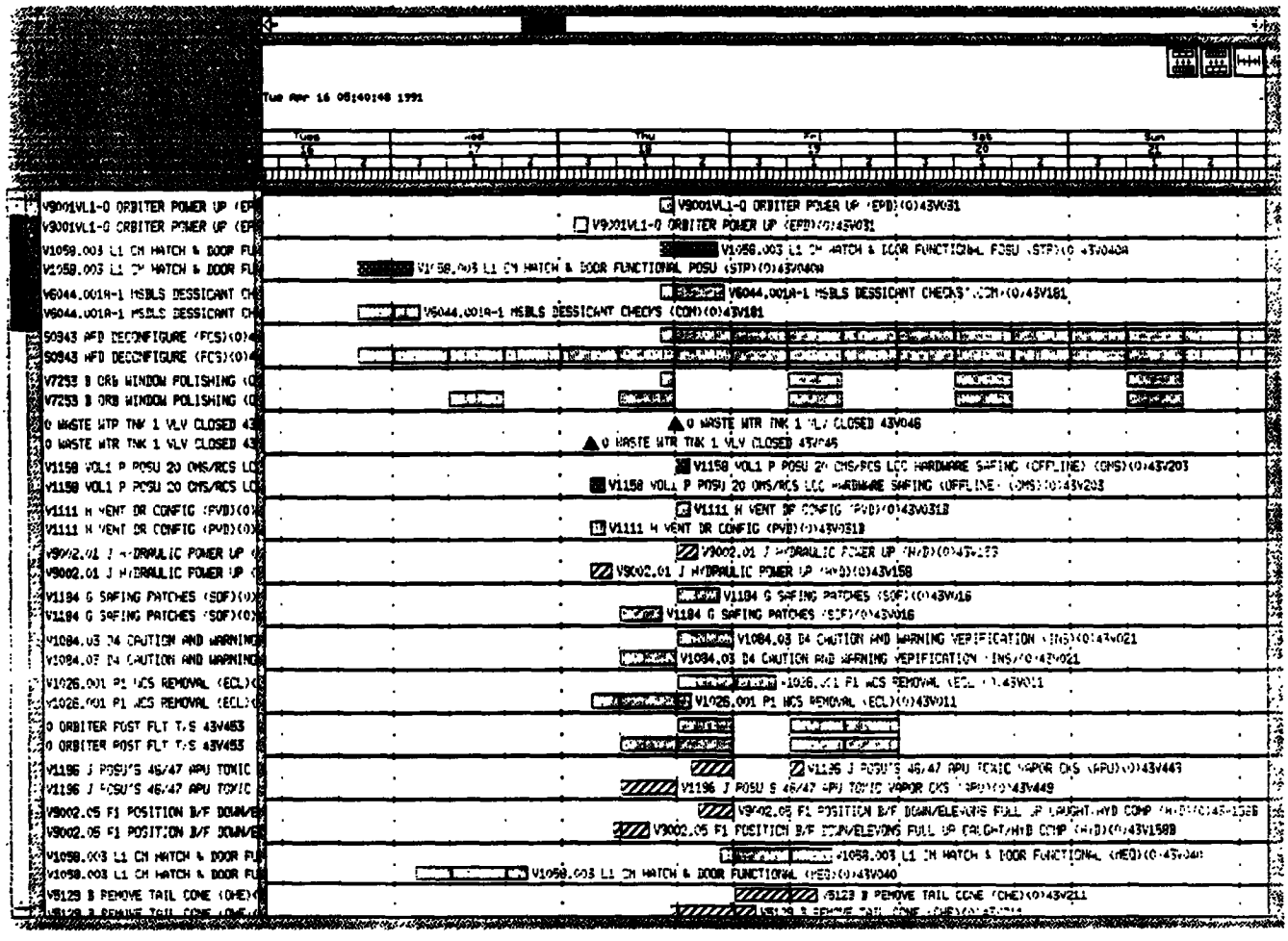


Figure 6.

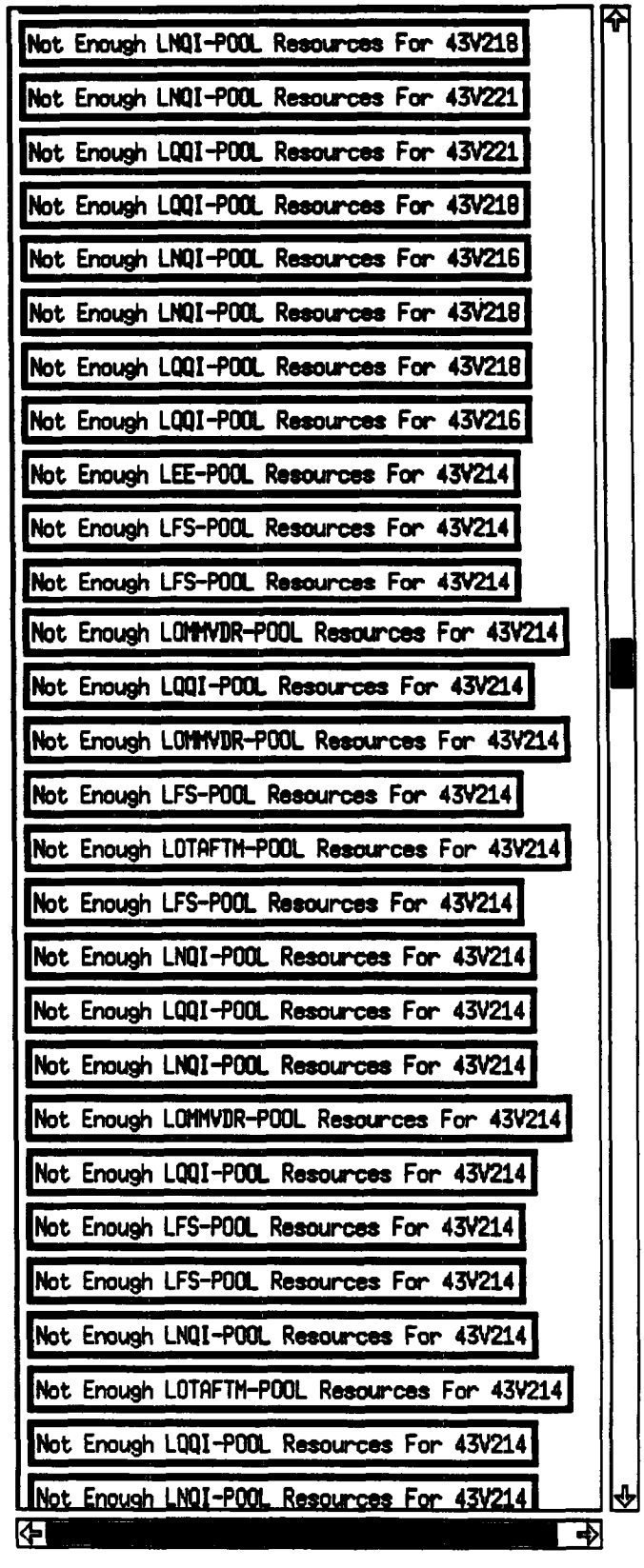

Not Enough LNQI-POOL Resources For 43V218
Not Enough LNQI-POOL Resources For 43V221
Not Enough LQQI-POOL Resources For 43V221
Not Enough LQQI-POOL Resources For 43V218
Not Enough LNQI-POOL Resources For 43V216
Not Enough LNQI-POOL Resources For 43V218
Not Enough LQQI-POOL Resources For 43V218
Not Enough LQQI-POOL Resources For 43V216
Not Enough LEE-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LOMMVDR-POOL Resources For 43V214
Not Enough LQQI-POOL Resources For 43V214
Not Enough LOMMVDR-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LOTAFTM-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LNQI-POOL Resources For 43V214
Not Enough LQQI-POOL Resources For 43V214
Not Enough LNQI-POOL Resources For 43V214
Not Enough LOMMVDR-POOL Resources For 43V214
Not Enough LQQI-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LFS-POOL Resources For 43V214
Not Enough LNQI-POOL Resources For 43V214
Not Enough LOTAFTM-POOL Resources For 43V214
Not Enough LQQI-POOL Resources For 43V214
Not Enough LNQI-POOL Resources For 43V214

Figure 7

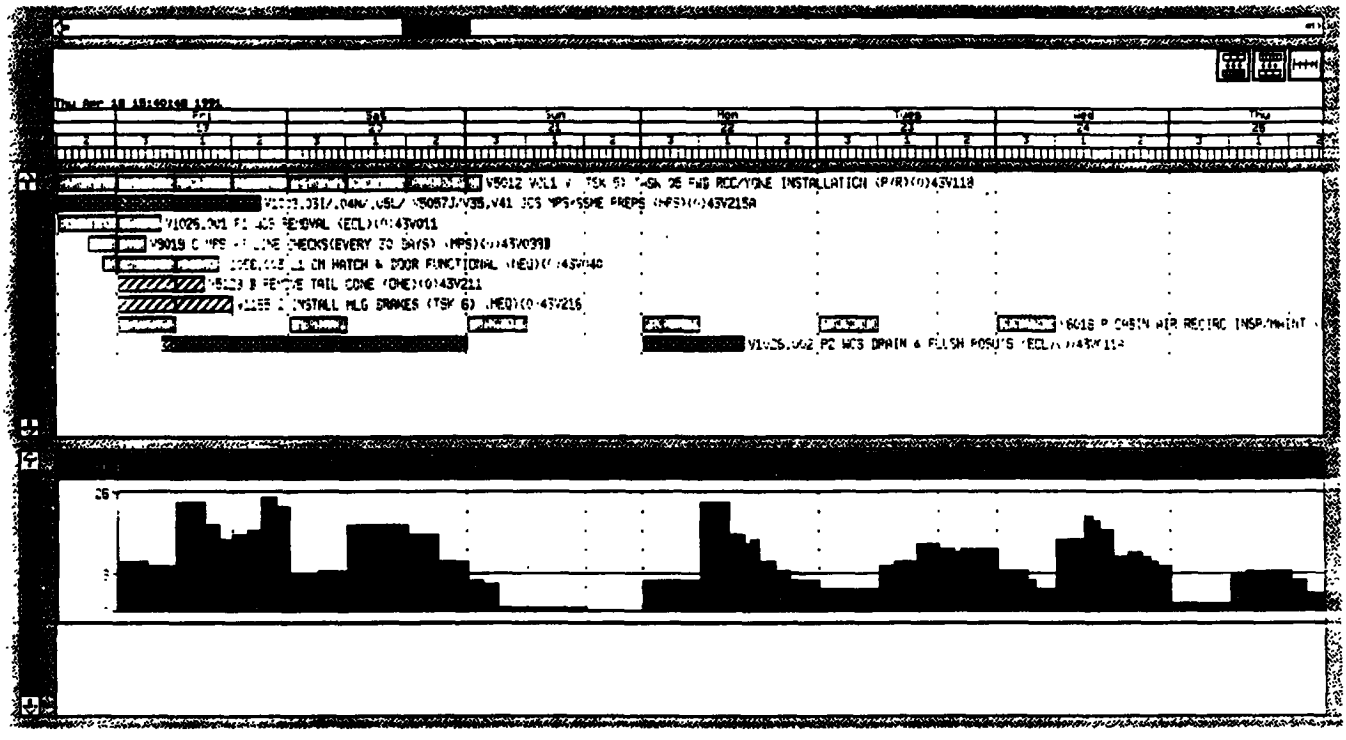


Figure 8.

ORIGINAL PAGE IS
OF POOR QUALITY

N92-22715

COMPASS: A GENERAL PURPOSE COMPUTER AIDED SCHEDULING TOOL

Mary Beth McMahon and Dr. Barry Fox
McDonnell Douglas Space Systems Company
16055 Space Center Boulevard
Houston, TX 77062

Chris Culbert
Software Technology Branch
NASA/Johnson Space Center - PT4
Houston, TX 77058

ABSTRACT

COMPASS is a generic scheduling system developed by McDonnell Douglas under the direction of the Software Technology Branch at NASA/Johnson Space Center. COMPASS is intended to illustrate the latest advances in scheduling technology and provide a basis from which custom scheduling systems can be built. COMPASS was written in Ada to promote readability and to conform to potential NASA Space Station Freedom standards. COMPASS has some unique characteristics that distinguishes it from commercial products. This paper discusses these characteristics and uses them to illustrate some differences between scheduling tools.

INTRODUCTION

Scheduling problems are wide-spread throughout NASA. Everything from experiments on-board the Space Station Freedom, to training facilities, to vehicle processing of the Space Shuttle require schedules to ensure effective use of resources and proper sequencing of activities. The construction of these schedules is often a labor intensive, time consuming activity involving interaction among many people. Many scheduling problems are seemingly simple; involving the placement of a defined set of activities on a timeline. However, considerable complexity is involved in keeping track of all the constraints that may inhibit a schedule and managing the effective use of resources needed to perform the activities. Manual systems have performed these jobs in the past with varying levels of efficiency. Computer aided solutions can significantly enhance performance in this domain by letting the computer track the large number of constraint details and allowing the human to focus on defining the basic structure of a schedule and resolving priority conflicts.

COMPASS is an automated scheduling system which can be used in a highly interactive manner. It allows a human scheduler to control the placement of activities on the schedule while COMPASS maintains the integrity of the schedule by tracking and enforcing resource and temporal constraints. It is generic in the sense that it was not developed for any specific problem domain. Rather, any scheduling problem that can be described by activities, resources, and their constraints can be modelled using COMPASS.

The motivation behind COMPASS was to provide advanced scheduling technology to the various NASA groups working on scheduling problems. It was written in Ada to promote readability and reusability. Anyone working on a government project may obtain a copy of the executable and the source code to read or modify as needed. It is very portable and currently runs on a variety of machines. To broaden the potential number of users, COMPASS supports both a graphical interface and a command line interface. The graphical interface is based on X-Windows and runs stand-alone on a SUN3/SUN4, SPARC station, IBM RS6000 and DEC VAX. In the client/server mode, it can be run over internet using an X-Window terminal to display the image while executing on any of the stand-alone machines. For those with minimal computing resources, a command line interface is available which prompts the user for commands. This interface has the same scheduling power as the graphical interface while retaining maximum portability.

This paper will describe some basic scheduling concepts and will explore traits of COMPASS that distinguish it from commercially available products. This knowledge is useful in analyzing scheduling problems and determining which tools and algorithms best fit a particular scheduling need. The traits discussed in this paper also serve as a measure by which to compare many of the commercial scheduling tools currently available.

DATA REPRESENTATIONS

There are four primary data objects that COMPASS uses to describe scheduling: Activities, Resources, Conditions and Time. Activities are the items which must be scheduled. Resources are the objects which must be present in order for an activity to be performed. These include such things as tools, electricity, and people. Conditions describe the state of the scheduling world. Activities may require that certain conditions exist before they are scheduled. Conditions may include things such as in-orbit, daytime, Phase III, etc. Time is used in all scheduling packages; however, the way it is represented affects the largest and smallest time units that may be used by the problem.

The Activity Data Structure

COMPASS provides thirteen fields to describe an activity. All but two are optional. This allows the user to create a data structure that closely fits the scheduling problem. The two mandatory fields are the activity name and the duration of the activity. The following is a list of all of the fields of the activity data structure.

Name	This is used as the unique id for the activity
Duration	The length of the activity (in time units)
Priority	The relative priority of the activity
Earliest Start	The earliest time the activity may start
Latest Finish	The latest time the activity may finish
Predecessors	List of activities that must precede this activity.
Successors	List of activities that must follow this activity.
Temporal Constraints	Timing constraints between the start or finish of one activity and the start or finish of another. There are four types of relations that may be represented: start/start, start/finish, finish/start, and finish/finish.
Nonconcurrent Activities	List of activity names that may not occur the same time as this activity.
Preferred Intervals	List of times during which it is preferred that this activity occur. These times must fall between the earliest start and latest finish.
Excluded Intervals	List of times between the earliest start and latest finish in which the activity may not occur.
Resource Requirements	Names of resources and their quantities required by this activity.
Condition Requirements	Names of conditions and the times which they must exist in order for this activity to be scheduled.

Most scheduling tools support at least some of the following fields: name, duration, priority, earliest start, latest finish, resource requirements, predecessors, successors and to some degree temporal constraints. The other fields are unique to COMPASS and have been included as optional fields which are can be used to help COMPASS fit actual scheduling problems. Due to the modularity in the design of the activity data structure, it is trivial to add new fields to the activity. The scheduling engine is easily modified to conform to the requirements of the new field.

The Resource Data Structure

A resource is typically represented by a resource name and a quantity for a given time period (eg. three men for two hours). These resource descriptions are typically used in three places.

(1) They are used to describe the initial availability of the resources. For example, initially there may be 5 men for available for 5 days.

(2) They are used to describe how much of a resource remains after activities have been scheduled. For example, after an activity is scheduled there may be only 4 men available for 5 days.

(3) They are used to describe the quantity of resources required by an activity. For example, an activity may require 1 man for 5 days.

There are several characteristics of resources that must be considered when describing a scheduling problems. Resource descriptions may be broken down into two types: Piecewise Constant and Piecewise Linear. Piecewise constant resource descriptions allow the use to request a quantity over an interval of time. More complex piecewise constant descriptions allow several quantities over several intervals of time. Piecewise linear descriptions specify a rate of consumption or production. An example might be the consumption of water - an activity might use a quart of water per minute for 1.5 hours. The closest approximation to this using piecewise constant is to describe it using "stair steps". However, this is more difficult to read and understand and is also not as accurate. A piecewise linear description would be a straight line whose second endpoint is 90 quarts lower and 1.5 hours later from its first endpoint.

Another important characteristic of resources is that they may be assignable, consumable or producible. An assignable resource is one that is used for the duration of the activity and then returned. A consumable resource is one that is consumed during the activity; therefore, when the activity finishes, less of that resource is available. A producible resource is one that is produced by an activity. It does not exist before the activity, but after the activity is under way the resource is available to be used by other activities. An example of an activity that both produces and consumes resources is the electrolysis of water. Water is consumed (at a rate) by the activity while hydrogen and oxygen are produced.

Most smaller (PC based) scheduling tools support only the assignable resources and the first instance of piecewise constant resource descriptions. Larger software products usually only support assignable resources and both instances of piecewise constant resource descriptions. COMPASS supports all of the above types of resources and resource descriptions.

The Condition Data Structure

A concept that is not commonly supported by scheduling tools is the ability to schedule activities only when certain conditions exist. COMPASS provides this capability using a data structure called Conditions. Conditions are used to describe the state of the world. Conditions are propositions whose values change between true, false and undefined over time. A condition is represented by a name and the intervals of time when the condition is true, false or undefined. If an activity requires that a certain condition be true, then the scheduling algorithm only schedules that activity during the span of time when the said condition is true.

The Time Data Structure

Most scheduling packages use calendars to describe the times when resources are available and dates to specify when activities should begin and end. Typically, the smallest unit of time is one minute and the largest unit of time one year. Dates can range from the late 1900's to the early 2000's.

COMPASS supports date representations and relative time representations. Relative time allows the user to specify time requirements in relative measures. For example, the earliest start for an activity might be two weeks. This means that the earliest start time for the activity is two weeks after "time zero", where time zero is an anchor which does not correspond to a particular date. This allows the user to schedule activities two weeks before launch or one

week after launch, without knowing exactly when launch is. Relative dates may be positive (eg. Launch plus one week) or negative (eg. Launch minus 3 days).

CHARACTERISTICS OF SCHEDULING TOOLS

There are three characteristics of a scheduling process which affect the way the schedule is built and the quality of the resulting schedule. These characteristics have minimal support, if any, in most commercial products; however, they are fully supported by COMPASS.

The first characteristic to look for is whether or not the scheduling tool is interactive. Can the user affect the placement of activities when the schedule is being built? With most commercial products, all of the data is entered and then the schedule is created automatically placing each activity on the schedule as early as possible. COMPASS allows the user to put activities on the schedule individually or in groups. They can be placed as early as possible or as late as possible or at any feasible time in between. Activities are easily unscheduled and rescheduled to respond to various scenarios.

COMPASS also supports incremental scheduling. This allows the user to build a schedule one activity at a time. It is not necessary to have all of the activities defined before starting to build a schedule. Once an activity is placed on the schedule, adding new activities does not interfere with those already on the schedule. That is, inserting an activity into the schedule does not change the starting times of any of the activities already on the schedule. This allows a user to make late additions to the schedule without affecting already scheduled activities. It also allows the user to place high priority items at preferred times with the guarantee that lower priority items will not move them from their slots.

Finally, COMPASS supports non chronological scheduling. Non chronological scheduling allows the user to select any time in the future to place an activity. In chronological scheduling, a user starts at a designated time, schedules any activities that can be scheduled at that time, then moves forward in time until an event occurs which allows other activities to be scheduled. This may be compared to planning a weeks worth of activities. The non chronological method would provide a calendar on which you could see the entire week ahead of you and write down which activities you wish to do in any order on the days you wish to accomplish them. The chronological approach would require that you start working on activities Monday morning and as you complete each activity, you pick a new one to start. You continue working on activities in this manner throughout the week until all are accomplished. You ca . possibly guess when you might get to an activity, but you cannot say in advance that "I will do this Friday at 9:00". Rather, you would have to wait until Friday to see what you are working on and when you expect to complete it in order to determine if you can do a specified activity at 9:00.

THE SCHEDULING ALGORITHM

Another distinguishing characteristic of COMPASS is the scheduling engine. Most schedulers are based on the Critical Path Method (CPM) algorithm. CPM is a chronological based scheduling algorithm which schedules each activity at its earliest starting time. This time is determined by examining the latest completion time of all of its predecessors. CPM does not take into account any of the resource constraints. The result is a schedule which finishes in the shortest amount of time, but makes no consideration of resource usage. This type of scheduling algorithm is good for estimations and loading studies. A loading study is used to determine the number of resources needed to complete a schedule by a given date. But with most scheduling problems, the number of resources are limited. Therefore, oversubscribing the resources produces an infeasible schedule. To alleviate the problem of oversubscription, most CPM tools also provide resource leveling. This technique shifts activities forward in time until the resources are available.

COMPASS does not use CPM. It uses a scheduling algorithm which takes into account both the temporal constraints and resource constraints. Given an activity, it determines all of the feasible times that the activity may be scheduled. Then the user provides some direction on the placement of the activity. This algorithm allows the user to select the order in which activities are placed on the schedule, as well as influence the placement of each as it is being scheduled. The feasible intervals of time in which an activity may be scheduled is determined by calculating all of the intervals of time that are feasible for each constraint and then taking the intersection. This gives the flexibility

of choosing which constraints to enforce. It also provides the user with information on all feasible times that the activity may be scheduled and allows the user to select the placement.

The scheduling algorithm used by COMPASS allows the user to constrain both the temporal and resource constraints simultaneously. Most commercial products will create a schedule abiding by the timing constraints and oversubscribing the resources where necessary. Then if the user wishes, resource leveling can be applied. There are two ways resources can be leveled. First the activities can be moved forward in time to the point where enough resources are available to satisfy the activity. This causes the successors of this activity to also be moved forward in time. Second, the quantities required by the activity can be reduced and the duration increased proportionally. This kind of resource leveling produces potentially undesirable results: it changes the quantities and durations of activities which, in some instances, would not produce feasible results; or it may push activities forward in time past their deadlines. If there are hard limits on both time and resources CPM and resource leveling will not produce feasible schedules, because on an oversubscribed problem it will violate either the timing or resource constraints.

Because of the way COMPASS is designed, it will never violate the timing or resource constraints. Instead, it will inform the user that the activity cannot be scheduled. One reason for taking this approach is that most space related scheduling problems have hard limits on both the time and resources.

CONCLUSIONS

COMPASS is unique in that it provides the user more control over the placement of activities than most commercial products do. This is due to its core scheduling algorithm which determines all of the feasible intervals of an activity then allows the user to affect the placement of it. COMPASS is also unique in that it will constrain both the resources and temporal relations simultaneously. This makes COMPASS an effective tool for addressing those scheduling problems which have a fixed time to complete a job with a fixed set of resources.

ACKNOWLEDGMENTS

The development of COMPASS and related scheduling research is supported by NASA under NAS9-17885.

REFERENCES

1. Fox, B. R., Mixed Initiative Scheduling, AAAI - Spring Symposium on AI in Scheduling, Stanford, CA, 1989.

2. Fox, B. R., Non-Chronological Scheduling, AAAI - Spring Symposium on AI in Scheduling, Stanford, CA, 1989.

DATA AND INFORMATION MANAGEMENT

(Session E2/Room B1)

Thursday December 5, 1991

- **ELAS: Powerful General-Purpose Image Processing Software**
- **TAE Plus: A NASA Tool for Building and Managing Graphical User Interfaces**
- **Instrumentation, Performance Visualization, and Debugging Tools for Multiprocessors**
- **Operations Automation Using Temporal Dependency Networks**

PRECEDING PAGE BLANK NOT FILMED

N92-22716

ELAS - A POWERFUL, GENERAL PURPOSE IMAGE PROCESSING PACKAGE

David Walters
Dr. Douglas Rickman
Science and Technology Laboratory
Stennis Space Center
SSC, MS 39529

ABSTRACT

ELAS is a software package which has been utilized as an image processing tool for more than a decade by Universities, State and Federal agencies and the private sector. It has been the source of several commercial packages. Now available on UNIX workstations it is a very powerful, flexible set of software. Applications at Stennis Space Center have included a very wide range, including medicine, forestry, geology, ecological modeling, and sonar imagery. It remains one of the most powerful image processing packages available, either commercially or in the public domain.

INTRODUCTION

ELAS was developed in the late 1970's by the Earth Resources Laboratory of the National Space Technology Laboratories. Known today as the Science and Technology Laboratory (STL) of Stennis Space Center, this organization is still involved in the development and application of ELAS software.

Originally created to process digital images acquired by the Landsat Multispectral Scanner, ELAS has developed into a very broad, general purpose raster processing tool. It has been used to process data from satellite and aircraft; images of Egyptian tomb paintings; fish scales and turtle flippers; MRI images of the human head, breast and heart; aerial photographs; soil maps; gravity potential fields; topographic data; and submarine sonar images. Areas of application have included forestry, agriculture, geology, archaeology, oceanography, medicine, ecology, environmental analysis, sonar imagery, and microclimatology.

DESCRIPTION

The ELAS software package is a modular approach to image processing. Predetermined processing runstreams are not provided. Image processing capability is broken down into components called application modules. These modules are considered building blocks which can be arranged by the user in an infinite variety of ways. The ELAS user has considerable control over how a module performs its task, as well as the order of execution. Each module allows the user to set the value of a number of processing parameters which define or control the module's operation.

Much of ELAS processing is performed on raster data which is stored in an ELAS specific format. This raster data may be images or any other type of data which could be stored as a two- or three-dimensional array. These data sets may be 8, 16, or 32 bit integer, 32 bit floating point, 64 bit floating point, complex or ASCII strings.

The package also has extensive ability to define and manipulate data defined in an x,y string format (polygons, line segments, and points). These are stored within the control file and allow a single numeric attribute, 0 - 255, to be attached to each vector. "Handedness", nodes, or other types of information are not retained.

PRECEDING PAGE BLANK NOT FILMED

ELAS consists of 239 applications modules. These modules can be categorized into one or more of the following functional groups:

FILE MANAGEMENT

This group consists of utilities to allocate, assign, release, copy, etc. various types of files used in ELAS.

REFORMATTING

Modules belonging to this group are used to import data into ELAS format files and to export ELAS files to other data formats. These foreign formats may consist of satellite, aircraft, vector, digital elevation, or various other raster and non-raster data.

DATA FILE UTILITY

These are modules that allow the user to manipulate the format, geometry, and contents of raster, and vector data files.

SUBFILE AND EXTERNAL FILE UTILITY

These modules are designed to build or manipulate various ELAS files which are not raster or vector data.

INTERACTIVE DATA DISPLAY

These modules are used for interactive work involving the display of images. The display medium may be a color display device or a window on a graphics terminal.

DESCRIPTIVE STATISTICS

This functional group includes routines which permit the user to examine specific characteristics of a data set without use of advanced statistical techniques.

STATISTICAL DATA ANALYSIS

These algorithms characterize a data file by some relatively sophisticated statistical measurement or property.

GENERATION OF TRAINING STATISTICS

This functional group consists of processes which gather statistical information to generate a definition of a class.

MANIPULATION OF TRAINING STATISTICS

These modules are designed to manipulate training statistics or display the relationships between selected classes.

CLASSIFIERS

These modules are used to assign a pixel to a class. It utilizes the class definitions generated by the training modules.

OPERATIONS ON CLASSIFIED DATA

This functional group contains a variety of operations requiring a classified data set to guide or define processing.

GEOMETRIC OPERATIONS

The modules in this functional group are designed to change the geometry of an image contained in an ELAS data file. The major use of these modules has been to correctly locate an image with respect to the globe.

CORRECTIONS AND CLEANUP FILTERS

These modules are used to correct both random and systematic errors in data or removal of noise and anomalies.

MODELING

This functional group contains modules to evaluate or model relationships within the data. The user may select either previously specified models or define his own.

POLYGON MANIPULATIONS

These modules give the capability to build, manipulate, and utilize vector data.

FILTERS

This functional group includes those modules that generate, use, or evaluate filters in either the time or frequency domain.

TOPOGRAPHIC DATA

This functional group includes modules which operate on digital elevation data.

PLOTTER AND FILM RECORDER

Modules that generate files for plotting to a special device or writing to film are in this group.

DIGITIZER

All modules directly linked to digitizing, either automatic or manual, are in this group.

SPECIAL PURPOSE AND MISCELLANEOUS

These modules do not conform to any of the previous functional groups.

STRUCTURE

Each application module is implemented as a separate program unit. This program unit consists of a main driver, UXMAIN, and a major subroutine containing the application specific code. The driver routine controls loading of the next module and maintains a common block of information necessary for execution of the modules.

Modules are swapped in and out of memory as a user executes them. Only one module is resident in memory at any given time during a single ELAS session. This minimizes the memory requirements based on program size. In addition, data files are read and processed a raster line at a time, reducing the memory requirements for file processing. Many ELAS modules can be executed with as little as 300 KB of memory.

Each ELAS session requires the user to allocate a "control file" or use a pre-existing one. A control file is used for each project to maintain continuity between multiple processing sessions. This control file keeps track of device and file assignments, and stores processing control information for each of the modules that have been used. Other ancillary information such as vectors, various look-up tables, and statistical data, also reside in the control file.

The ELAS data file is the only file structure in ELAS for storing raster data. As mentioned earlier, data may be in any of several numeric representations. The ELAS file can contain any data which can be expressed in a two- or three-dimensional array. This is typically image data, but the data are not restricted to this. Most application modules are designed to process the ELAS files regardless of the contained data type. In combination these attributes allow the user to largely ignore limitations normally imposed in image processing for reasons of limited range in numeric representation.

OPERATION

As noted, on entry to ELAS the user is prompted to supply a control file name. During a single session all operations and control will be within this single environment. After selecting files for input and output, display and input devices the user may go to any of the modules. Once in the module a list of "directives" are available. Directives are major processing options within the program. One of the standard directives is an option to set parameters. Parameters are variables which control such things as regions of the input or output file to be processed, or number of classes to produce.

At any time the user may leave a module and go directly to any other module. The only exception is when the computer is executing a previous command. All program control information (parameters) will have been written into the appropriate portions of the control file, thus the user can return to any previously used module and continue processing where ever execution was finished. This also permits the user to determine the conditions under which file was manipulated. This ability to return also extends to abnormal terminations, such as system crashes. Almost all operations can be restarted with loss only of the last scan line of imagery needing to be rerun.

The user can also switch processing control to an ASCII file containing commands. Built by any ASCII capable word processor or system editor, these files permit the user to create processing runs consisting of thousands of commands. This facility realistically allows the knowledgeable operator to treat whole modules as building blocks, effectively creating new capabilities.

The documentation for each module is also on-line. This is the equivalent of a manual 3 1/4" thick. Documentation gives the purpose of the module, operating instructions, parameter settings, functions of directives, file and program restrictions, resources used and formats, and one or more examples.

HOST SYSTEMS

At this time, four versions of ELAS have been implemented on UNIX platforms by STL. The source code for the application modules are identical for each of the versions. The major difference in these UNIX versions is the interface to the graphical display device for image visualization. The following versions are currently supported by STL:

Masscomp

This version utilizes a separate GA1000 display terminal with the graphics calls implemented through Masscomp's GP library.

Sun

This version uses X11 and the XView toolkit to implement a display window. This requirement may be met with MIT X11.4 or the Sun OpenWindows Version 2. The Open Look Window Manager is required. Currently, only 8-bit color look-up display is available in this version.

Silcon Graphics

This version uses the Silicon Graphics specific window system, Foresight, to implement display windows. This also requires use of the GL graphics library. This version was developed for a machine with the Super Graphics giving true-color capability and an extended color look-up table. Multiple ELAS tasks can be executed, each with an independent 8 bit pseudo-color display or a true color (24 bits - 8 bit each RGB) display. All displays will also have multiple independent graphic overlays. Multiple displays on the other UNIX window versions share a common 256 color look-up table.

Data General

This version uses X11 and Xt libraries to implement a display window. The Athena widget set is required. As with the SUN version, this is an 8 bit implementation for display.

The above versions are all public domain and can be obtained from:

COSMIC
University of Georgia
Athens, Georgia 30602
(404) 542-3265

Other, non-UNIX versions of ELAS have been created in the past for a large number of machines. Included are the Concurrent, VAX, Prime, SEL, Varian. ELAS has also been used as the root of commercial packages, such as ERDAS and ATLAS.

CONCLUSION

ELAS is an extremely powerful image processing package. It is well suited to applications which need a very large degree of flexibility. Indeed, one of the programming axioms is to not prohibit the user from gaining access or doing anything, unless it directly violates the basic algorithm or mathematics. This freedom means power; it also means the user has no single "yellow brick road". The software is command line driven, not menu oriented. These characteristics mean that an experienced user can make the software do more, do it faster, and in more ways. However, this requires the novice to invest some time in training.

With the availability of commercial spin offs and image processing software from many other sources, much of ELAS is no longer unique. There remain however several areas in which ELAS is not equalled or excelled. Filtering, statistical analysis, geometric correction, high degree of user control over algorithm execution, and flexibility gained through modularity are still major strengths.

Because of the nature of the software, ELAS has always been regarded as a significant burden for most machines. The workload of its algorithms have outweighed the power of the cpu and I/O capacity. The development of low-cost, high-performance workstations during the past few years, may help to overcome this problem and allow ELAS to reach its true potential. With the availability of ELAS on UNIX workstation platforms such as SUN, MassComp, Silicon Graphics, and Data General, the popularity of ELAS will continue to grow.

N92-22717

TRANSPORTABLE APPLICATIONS ENVIRONMENT (TAE) PLUS

a NASA tool for Building and Managing Graphical User Interfaces

Martha R. Szczur
NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
mszczur@postman.gsfc.nasa.gov
301 286-8609

ABSTRACT

The Transportable Applications Environment (TAE) Plus, developed at NASA's Goddard Space Flight Center, is an advanced portable user interface development environment which simplifies the process of creating and managing complex application graphical user interfaces (GUIs), supports prototyping, allows applications to be ported easily between different platforms and encourages appropriate levels of user interface consistency between applications. This paper will discuss the capabilities of the TAE Plus tool, and how it makes the job of designing and developing GUIs easier for the application developers. TAE Plus is being applied to many types of applications, and this paper discusses what TAE Plus provides, how the implementation has utilized state-of-the-art technologies within graphic workstations, and how it has been used both within and outside NASA.

BACKGROUND

Emergence of graphical user interfaces

With the recent emergence of sophisticated graphic workstations and the subsequent demands for highly interactive systems, designing and developing good user interfaces has become more complex and difficult. Prior to the graphic workstations, the application developer was primarily concerned with developing user interfaces for a single monochrome 80x24 alphanumeric character screen with keyboard user entry. With high resolution bit-mapped workstations, the user interface designer has to be cognizant of multiple window displays, the use of color, graphical objects and icons, and various user selection techniques (e.g., mouse, trackball, tablets).

High resolution graphic workstations also provide system developers with the opportunity to rethink and redesign the user interfaces (UI) of their next generation applications. For instance, in a command and control environment, many processes run simultaneously to monitor a particular operation. With modern graphic workstations, time-critical information concerning multiple events can be displayed concurrently on the same screen, organized into different windows in a variety of graphical and textual presentations. As today's workstations inspire more elaborate user interfaces, the applications which utilize their graphics capabilities increase in complexity. Productivity tools to aid in the definition and management of user interfaces, thus, become an increasingly important element in the application's prototyping-to-operational development cycle.

Requirements for a prototyping-to-operational development environment

To support our development cycle we wanted to establish an integrated environment that allows prototyped user interfaces to evolve into operational applications. This environment would satisfy the following objectives:

- separate the user interface from the application,
- provide tools to allow interactive design/change/save of user interface elements,
- take advantage of the latest hardware technology,
- support rapid prototyping,
- manage the user interface,
- develop tools for increasing application development productivity,
- provide the application with runtime services, and
- allow portability to different computing environments.

Building on existing technology

Many of these objectives were addressed in the early 1980's when GSFC recognized that most large-scale space applications, regardless of function, required software to support human-computer interactions and application management. This lead to the design and implementation of the Transportable Applications Executive (now, referred to as TAE *Classic*), which abstracts a common core of system service routines and user dialog techniques used by all applications[1]. Over the years, TAE Classic has matured into a powerful tool for quickly and easily building and managing consistent, portable user interfaces, but only for the standard alphanumeric terminal. Not only did TAE Classic improve the productivity of a single application's development life cycle by providing the programmers with an easy and standard method for creating menus, prompting for parameters and building command procedures, but, because the tool was generic and reusable for multiple applications, the productivity gain for implementation increased exponentially. Other gains were realized in a significant reduction in application testing time (i.e., the user interface component, TAE, is reliable, debugged software) and maintenance overhead (i.e., application code uses TAE services, thus becoming hardware and operating system independent, which simplifies making application changes and enhancements.)

In the past six years, the emergence of the low-cost graphic workstation has enabled development of innovative graphical user interfaces (GUIs). Along with this new capability and flexibility comes a significant increase in the complexity of developing these graphical user interfaces. The array of new UI elements associated with GUIs (e.g., windows, panes, color, direct manipulation, programmable cursors) requires the developer/programmer to understand a complex new software environment (e.g., the X Window System™, "widget" architecture, OSF/*Motif*™ and AT&T's Open Look™ user interfaces.) This required expertise can translate into an increase in programmer training. Further, maintenance nightmares can occur as the low-level windowing systems are upgraded/changed. Frequently the cost of the GUI development can increase to the point where it exceeds the application-specific components. At GSFC we wanted to take advantage of the new GUI capabilities, but needed a way to improve the productivity of developing an application's graphical user interface component. We took advantage of the lessons learned in the TAE Classic development. By utilizing some of the internal data structures and features of the original TAE software, we developed a set of tools which support the building and management of GUIs. This advanced version of TAE is called TAE Plus (i.e., TAE *Plus* graphics support).

WHAT DOES TAE PLUS PROVIDE?

To meet the defined goals, services and tools were developed for creating and managing window-oriented user interfaces. It became apparent, due to the flexibility and complexity of graphical user interfaces, that the design of the user interface should be considered a separate activity from the application program design. The interface designer can then incorporate human factors and graphic art techniques into the user interface design. The application programmer needs only to be concerned about what results are returned by the user interaction and not the look of the user interface.

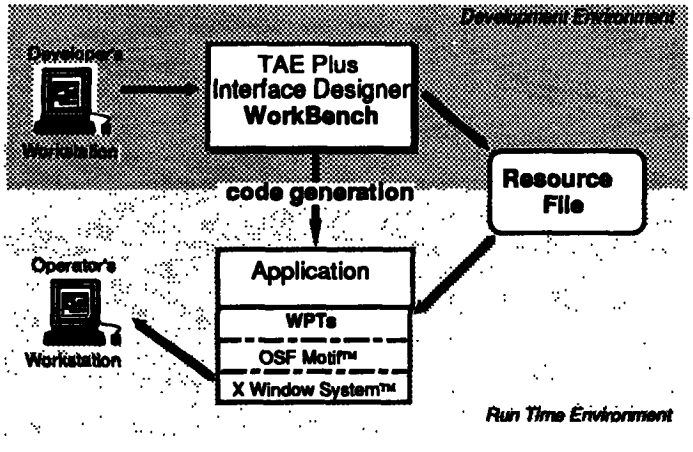


Figure 1. TAE Structure

In support of the user interface designer, an interactive *WorkBench* application was implemented for manipulating interaction objects ranging from simple buttons to complex multi-object panels. As illustrated in Figure 1, after designing the screen display, the WorkBench saves the specification of the user interface in resource files, which can then be accessed by application programmers through a set of runtime services, Window Programming Tools (WPTs). Guided by the information in the resource files, the routines handle all user interactions. The WPTs utilize Open Software Foun-

dation's Motif™ and the standard MIT X Window System™ to communicate with the graphic workstations.[2] As a further aid to the UI developer, the WorkBench provides an option to generate the source code which will display and manage the designed user interface. This gives the programmer a working template into which application-specific code can be added.

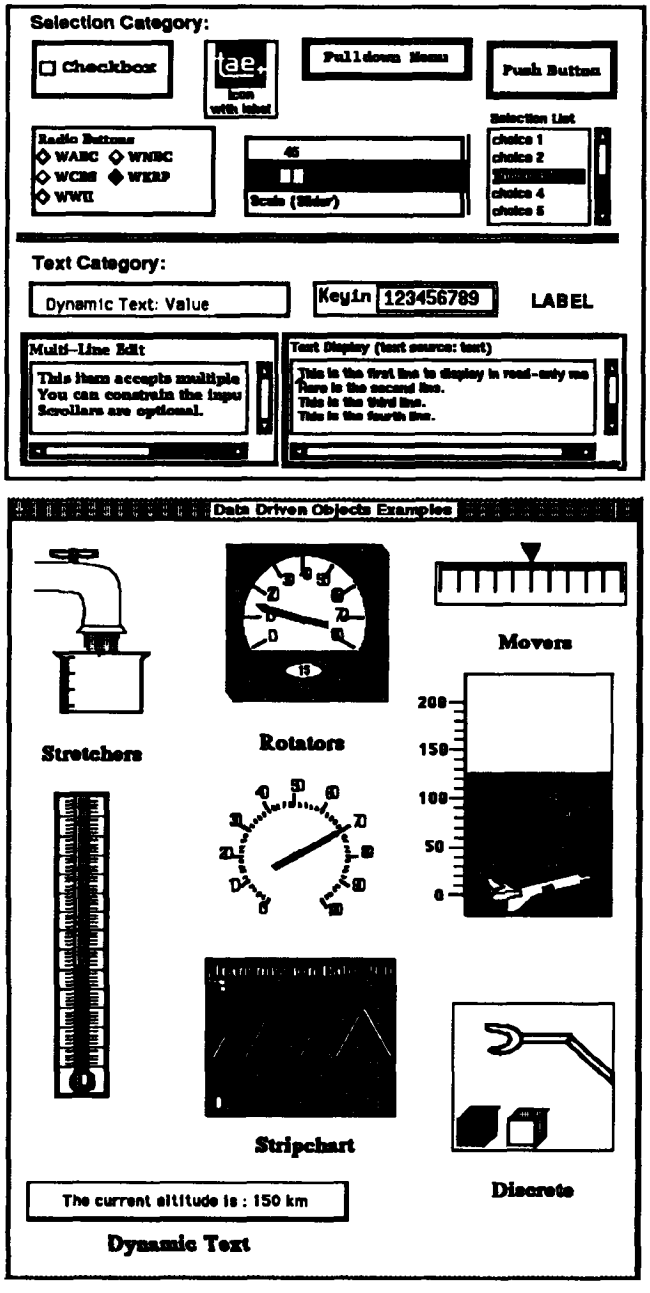


Figure 2. TAE Plus User Interface Interaction Objects

INTERACTION OBJECTS AS BUILDING BLOCKS

The basic building blocks for developing an application's GUI are a set of interaction objects. All visually distinct elements of a display that are created and managed using TAE Plus are considered to be interaction objects and they fall into three categories: user-entry objects, information objects, and data-driven objects. *User-entry objects* are mechanisms by which an application can acquire information and directives from the end user. They include radio buttons, check boxes, text entry fields, scrolling text lists, pulldown menus and push buttons. *Information objects* are used by an application to instruct or notify the user, such as contextual on-line help information displayed in a scrollable static text object or brief status error messages displayed in a bother box. *Data-driven objects* are vector-drawn graphic objects which are linked to an application data variable; elements of their view change as the data values change. Examples are dials, thermometers, and strip charts. When creating user dialogues, these objects are grouped and arranged within *panels* (i.e., windows) in the WorkBench.

The use of interaction objects offers the application designer/programmer a number of benefits with the expected payoff of an increase in programmer productivity. The interaction objects provide a consistent look and feel for the application's user interface, which translates into reduced end-user training time, more attractive screens, and an application which is easier to use. Another key benefit is that since the interaction objects have been thoroughly tested and debugged, the programmer is able to spend more time testing the application and less time verifying that the user interface behaves correctly. This is particularly important considering the complexity of some of the objects, and the programming effort it would take to code them from scratch. Refer to Figure 2 for a sample of the TAE Plus interaction objects.

TAE PLUS WORKBENCH

The WorkBench provides an intuitive environment for defining, testing, and communicating the look and feel of an application system. Functionally, the WorkBench allows an application designer to dynamically lay out an application screen, defining its static and dynamic areas. The tool provides the designer with a choice of pre-designed interaction objects and allows for tailoring, combining and rearranging of the objects. To begin the session, the designer needs to create the base panel (i.e., window) into which interaction objects will be specified. The designer specifies presentation information, such as the title, font, color, and optional on-line help for the panel being created. The designer defines both the presentation information and the context information of all interaction items to reside in the panel by using the item specification window (refer to Figure 3). For icon support, the WorkBench has an icon editor, within which an icon can be drawn, edited and saved. As the UI designer moves, resizes, and alters any of the item's attributes, the changes are dynamically reflected on the display screen.

The designer also has the option of retrieving *palettes* of previously created items. The ability to reuse interaction objects saves programming time, facilitates experimenting with different combinations of items in the prototyping process, and contributes to standardization of the application's look and feel. If an application system manager wants to ensure consistency and uniformity across an entire application's UI, all developers could be instructed to use only items from the application's palette of common items.

When creating a data-driven object, the designer goes through a similar process by setting the associated attributes (e.g., color thresholds, maximum, minimum, update delta) in the specification panels. To create the associated graphics drawing, the WorkBench provides a drawing tool within which the static background and dynamic foreground of a data-driven object can be drawn, edited, and saved. Figure 4 shows the drawing tool being used to create a *stretcher* data-driven object.

Most often an application's UI will be made up of a number of related panels, sequenced in a meaningful fashion. Through the WorkBench, the designer defines the interface *connections*. These links determine what happens when the user selects a button or a menu entry. The designer attaches *events* to interaction items and thereby designates what panel appears and/or what program executes when an event is triggered. Events are triggered by user-controlled I/O peripherals (e.g., point and click devices or keyboard input).

TAE Plus also offers an optional help feature which provides a consistent mechanism for supplying application-specific information about a panel and any interaction items within the panel. In a typical session, the designer elects to edit a help file after all the panel items have been designed. Clicking on the edit help option in the Panel Specification Panel brings up a text editor window in which the appropriate information can be entered. The designer can then define any button item or icon item to be the help item for the panel (in this scenario it would be the help icon in the panel "Monitor".) During the application operation, when the end-user clicks on the question mark item, the cursor changes to a question mark symbol (?). The end-user then clicks on the panel itself or any item in the panel to bring up a help panel containing the associated help text.

Having designed the layout of panels and their attendant items and having threaded the panel and items according to their interaction scenario, the designer is able to preview (i.e., rehearse) the interface's operation from the Work-Bench. With this potential to test drive an interface, to make changes, and to test again, iterative design becomes part of the development process. With the rehearsal feature, the designer can evaluate and refine both the functionality and the aesthetics of a proposed interface. After the rehearsal, control is returned to wherever the designer left off in the WorkBench and the designer can either continue with the design process or save the defined UI in a resource file.

Developing software with sophisticated user interfaces is a complex process, mandating the support of varied talents, including human factors experts and application program specialists. Once the UI designer (who may have limited experience with actual code development) has finished the UI, he/she can turn the saved UI resource file over to an experienced programmer. As a further aid to the application programmer, the WorkBench has a "generate" feature, which produces a fully annotated and operational body of code which will display and manage the entire WorkBench-designed UI. Currently, source code generation of C, Ada, and the TAE Command

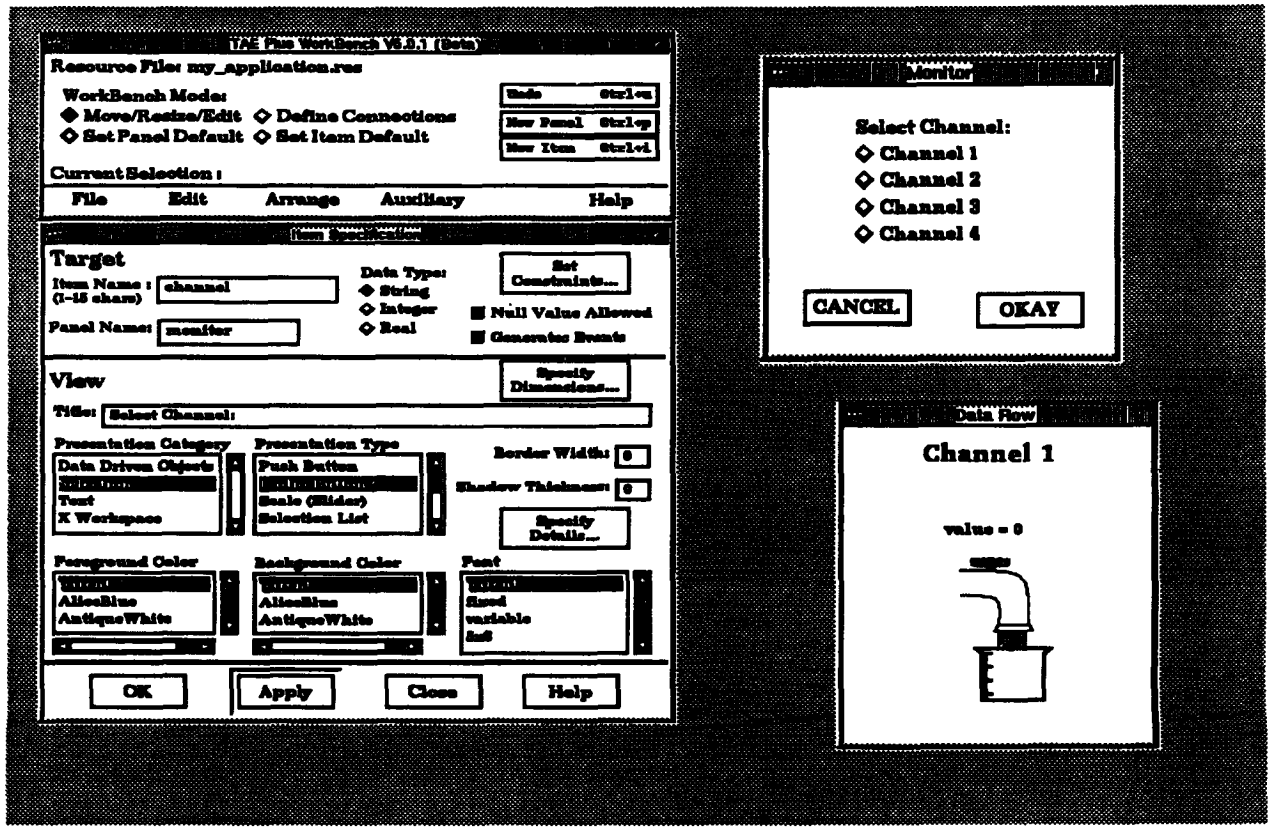


Figure 3. Building a user interface with the WorkBench

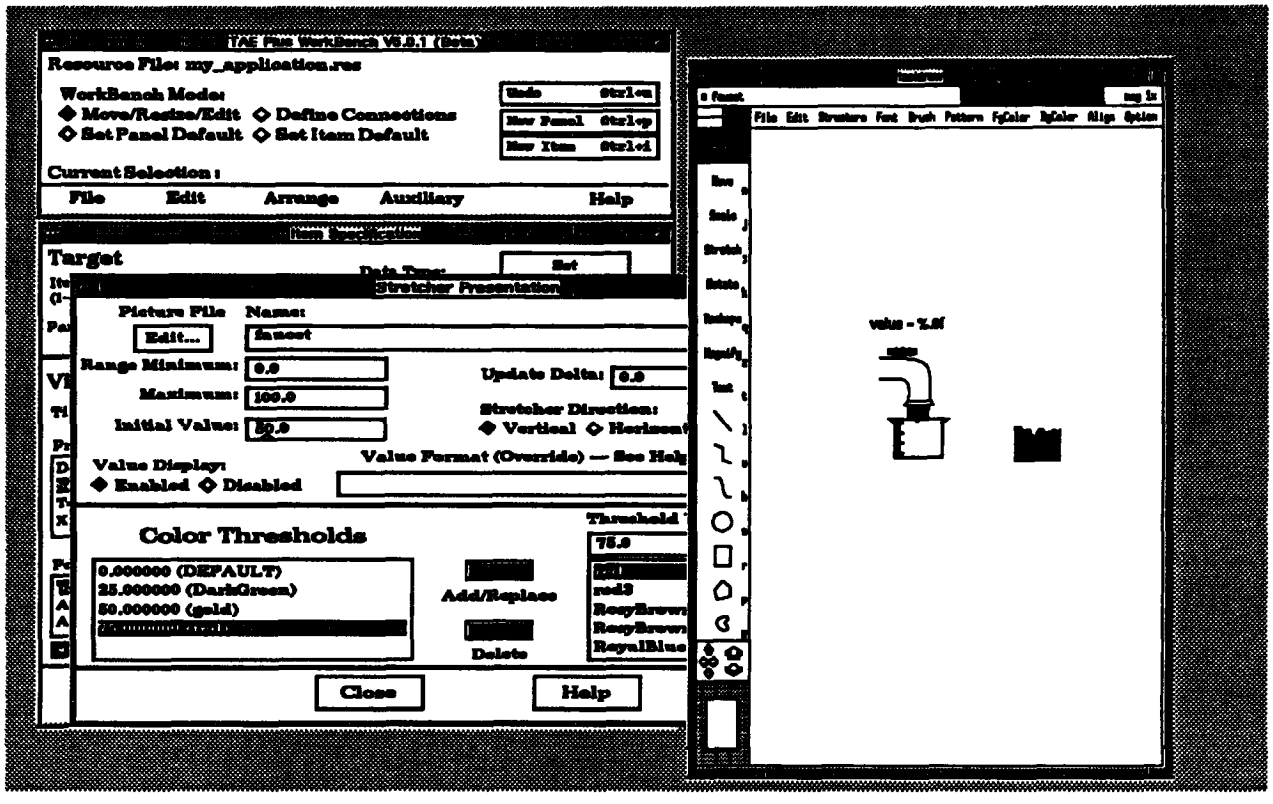


Figure 4. Creating a stretcher data-driven object

Language (TCL) (an interpreted prototyping language) are supported, with bindings for C++ expected in a future release of TAE Plus. The programmer can now add additional code to this template and make a fully functional application. Providing these code stubs helps in establishing uniform programming method and style across large applications or within a family of interrelated software applications.

WINDOW PROGRAMMING TOOLS (WPTS)

The Window Programming Tools (WPTs) are a package of application program callable subroutines used to control an application's user interface. Using these routines, applications can define, display, receive information from, update and/or delete TAE Plus panels and interaction objects. WPTs support a modeless user interface, meaning a user can interact with one of a number of interaction objects within any one of a number of displayed panels. In contrast to sequential mode-oriented programming, modeless programming accepts, at any instance, a number of user inputs, or *events*. Because these multiple events must be handled by the application program, event-driven programming can be more complex than traditional programming. The WorkBench's auto-generation of the WPT event loop reduces the risk of programmer error within the UI portion of an application's implementation.

The WPT package utilizes the the MIT X Window System, as its standard windowing system. One of the strengths of X is the concept of providing a low-level abstraction of windowing support (Xlib), which becomes the base standard, and a high-level abstraction (X toolkits), which has a set of interaction objects (called "widgets" in the X world) that define elements of a UI's look and feel. Due to the growing acceptance of the OSF/ Motif user interface style as a defacto industry standard, the latest release of TAE Plus (V5.1) is based on the Motif software.

The WPTs also provide a buffer between the application program and the Motif toolkit and Xlib services. For instance, to display a WorkBench-designed panel, an application makes a single call to Wpt_NewPanel (using the *panel name* specified in the WorkBench). This single call translates into a function that can make as many as 50 calls to Motif library routines. For the majority of applications, the WPT services and objects supported by the WorkBench provide the necessary user interface tools and save the programmer from having to learn the complexities of programming directly with Motif and X. This can be a significant advantage, especially when considering the learning curve differential between 40 WPT routines versus over 400 X Toolkit intrinsics and over 200 Xlib services.

IMPLEMENTATION

The TAE Plus architecture is based on a separation of the user interaction management from the application-specific software. The current implementation is a result of having gone through several prototyped and beta versions of a WorkBench and user interface support services during the 1986-89 period, as well as building on the TAE Classic structure.

The "Classic" portion of the TAE Plus code is implemented in the C programming language. In selecting a language for the WorkBench and the WPT runtime services, we felt a "true" object-oriented language would provide us with the optimum environment for implementing the TAE Plus graphical user interface capabilities. (See Chapter 9 of Cox[4] for a discussion on the suitability of object-oriented languages for graphical user interfaces.) We selected C++[5] as our implementation language for several reasons[6]. For one, C++ is becoming increasingly popular within the object-oriented programming community. Another strong argument for using C++ was the availability of existing, public domain, X-based object class libraries. Utilizing an existing object library is not only a cost saver, but also serves as a learning tool, both for object-oriented programming and for C++. Delivered with the X Window System is the *InterViews* C++ class library and a drawing utility, *idraw*, both of which were developed at Stanford University[7]. The *idraw* utility is a drawing editor, which we integrated into the WorkBench to support creating, editing and saving the graphical data-driven interaction objects. This reuse of existing software enabled the addition of a major new function without the significant cost and time of implementing a drawing editor from scratch.

The single most important factor contributing to the portability of TAE Plus is the X Window System. Generally, if a graphic workstation supports the Xlib and the OSF/Motif X Toolkit and operates either UNIX or VMS, TAE Plus can be ported to it with reasonable ease. For instance, TAE Plus is operational on the following UNIX platforms: Sun workstations, Apollo, VAXstation II , DECstation 3100, HP9000, Masscomp, Silicon Graphics Iris, NEC EWS 4800/220 and Macintosh II (A/UX). TAE Plus is also available and validated on the VAXstation II and VAXstation 3100 under VMS.

TAE PLUS AS A PRODUCTIVITY TOOL

For years the software industry has been searching for ways to quantify the software development process allowing for accurate measurement of productivity. Due to the cerebral versus mechanical nature of software development this is a difficult task, which has lead to a large volume of published approaches on how to improve software productivity.[8] Barry Boehm identifies six primary options for improving software productivity[9] and TAE Plus addresses each one of these options at some level.

Getting the best from people

To get the maximum productivity from each member of a development team individuals should be utilized in the areas that they have an expertise. Too often the people designing application user interfaces are the programmers, who most often do not have any training in human factors or graphic art techniques. This tends to be an ineffective use of the programmers expertise, and frequently results in a less than optimum user interface. The WorkBench was designed to eliminate this problem giving the user interface design experts a tool that is easy to use (i.e., does not require programming skills), while freeing up the programmer to concentrate on the application specific code.

Make steps more efficient

As stated by Boehm, "the primary leverage factor in making the existing software process steps more efficient is the use of software tools to automate the current repetitive and labor-intensive portions of each step." Prior to a tool like the WorkBench, the layout of the user interface involved either paper and pen mockups and layouts, or programmers creating the UI as they coded. In either case, the availability of an interactive user interface layout tool that allows the designer to define and build the UI in a WYSIWYG manner, makes the UI design process more efficient.

Eliminating Steps

The next productivity option is to automate a previous manual step, thus eliminating the step entirely. TAE Plus provides the capability to automatically generate the application code that manages the designed UI. This eliminates the process of the application programmer having to manually generate and key in this code, thus reducing the likelihood of keyboard errors or incorrect function calls. Particularly in cases where the application is heavily interactive, this automatic code generation can account for the majority of the application code and significantly improve productivity of the development process.

Eliminating Rework

Information hiding and prototyping are both ways that contribute to avoidance of reworking code. In TAE Plus, all the details of the user interface are hidden from the application. For instance, the application calls on a WPT routine (i.e., Wpt_NewPanel) to display a panel and its interaction objects. The application is not interested in whether the user is being presented with a radio button bank or a scrollable text list, but it is interested in which choice the user makes from this interaction object. The actual display and management of the UI is handled within the WPTs, thus isolating the UI code from the application. During an application's evolution, this approach of hiding details within the WPTs minimizes or eliminates the impact that changes to the UI hardware, the windowing system, the object class, etc., will have on the application.

Building simpler products

The number of software source instructions programmed during an application's development has the most significant influence on software costs. One approach to improving productivity is to reduce this number by building simpler products and eliminating "software gold plating: extra software that not only consumes extra effort but also reduces the conceptual integrity of the product."[9] Using rapid prototyping as a step in the specification process can frequently prevent the over specification of functions by users who are worried that if they don't specify everything

that can think of, then the system will not have some function they need. Although there is no guarantee that rapid prototyping will result in a simpler program, it fosters a dialog between the developers and the user that can solidify the real system requirements and specifications. As a tool that enables rapid prototypes to be built quickly and easily, TAE Plus can be used to design simpler applications.

Reusing Components

Another way to reduce the amount of source code written for an application is to reuse existing software. TAE Plus was designed with software reuse as a primary goal. The WPT runtime services offload all of the display and management of the UI from the application code. This approach enables the application programmer to concentrate fully on the application-specific functions, and not be concerned with the UI code. Also, TAE Plus itself reuses existing standard windowing software (e.g., MIT's X Window System, OSF/Motif, Stanford's Interview object classes), thus improving the productivity of its own development.

TAE PLUS CASE STUDIES

One way to measure how effective TAE Plus is as a productivity tool is to develop the same application twice, one time using TAE Plus and another time not using TAE Plus. While most users feel certain that TAE Plus is saving them development time, they are on tight development schedules and do not have the interest in building parallel UIs. However, a few case studies in which the same user interface was developed with and without TAE Plus give evidence that the productivity gain can be impressive.

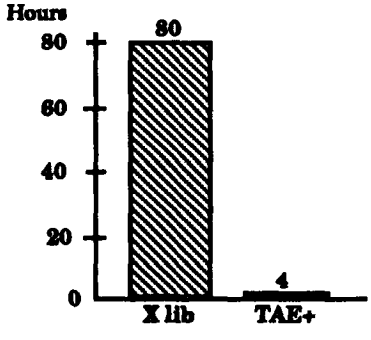


Figure 6. Case Study 1

In Case 1, a programmer from General Electric developed a simple screen copy utility which gathers information through radio buttons, action icons, and text input. Then, it sends the information to an HP printer, as well as updating a text widget on the screen. When he did not use TAE Plus and wrote the UI code directly within the application code, it took him 80 hours to develop an operational application. When he used the TAE Plus WorkBench to develop the same operational application, it took him 4 hours. This productivity gain of 95% is illustrated in Figure 6. However, it should be noted that the gain does not take into account the unmeasured factor that "it is always easier the second time around."

Figure 2 illustrates Case 2. A programmer at NASA with no TAE Plus experience, but with X Window System experience, was tasked to write a simple application and account for the time spent on developing it with and without TAE Plus. The application has two panels, a few action icons, a radio button bank, and a dynamic mover object that moves along a static background when the associated data value changes. Including the time it took to learn how to use the WorkBench to the completion of the operational application, it took him 9 hours. (Note: an experienced TAE Plus user did the same application in 1.5 hours.) The application developed without TAE Plus (thus, making direct calls to the X Window System) took him 52 hours, and this implementation was still a "bit buggy." Even as a beginner TAE Plus user, it took him over four times longer to develop the application without TAE Plus. In the case of the experienced TAE Plus user, the productivity gain was even more dramatic, with a 96% increase in development of the application. Although these case studies certainly do not provide enough statistical data to allow any grandiose conclusions to be made, they do demonstrate real cases in which using a GUI development tool, in this case TAE Plus, has significantly decreased the time it takes to develop the application. In general, TAE Plus reduces the time it takes a developer to create, test and deliver a software system.

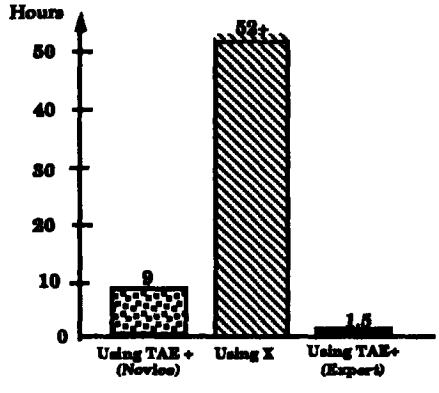


Figure 7. Case Study 2

AVAILABILITY AND MAINTENANCE

In April 1991 TAE Plus 5.1, which uses the latest version of OSF Motif™ (V1.1), became available from COSMIC, the NASA's software distribution center located at the University of Georgia. Versions for numerous UNIX workstations (e.g., Suns, DECstation 3100, HP9000, Apollo) and for VMS/DECWindows™ may be licensed at a nominal fee.

Maintenance of a software system is a key factor in its success, and while every system is maintainable, *how easy it is to maintain* is the real issue. We knew when we began development that TAE Plus was targeted for wide application utilization and for different machines, so ease of maintenance has always been important. By providing the application-callable WPTs, applications are isolated from the windowing system. Thus, when the latest release or next generation windowing system shows up, only the WPTs will require updating or rewriting; the application code will not be affected.

User support is another facet of maintainability. Since the first release of TAE Classic in 1981, we have provided user support through a fully staffed Support Office. This service has been one of the primary reasons for the success of TAE. Through the Support Office, users receive answers to technical questions, report problems, and make suggestions for improvements. In turn, the Support Office keeps users up-to-date on new releases, provides a newsletter, and sponsors user workshops and conferences. This exchange of information enables the Project Office to keep the TAE software and documentation "in working order" and, perhaps most importantly, take advantage of user feedback to help direct our future development.

APPLICATIONS USING TAE PLUS

Since 1982 over 900 installation sites have received TAE Classic and/or TAE Plus. The applications built or being built with TAE perform a variety of different functions. TAE Classic usage was primarily used for building and managing large scientific data analysis and database systems (e.g., NASA's Land Analysis System (LAS), Atmospheric and Oceanographic Information Processing System (AOIPS), and JPL's Multimission Image Processing Laboratory (MIPL) system.) Within the NASA community, TAE Plus is also used for scientific analysis applications, but the heaviest concentration of user applications has shifted to support of realtime control and processing applications. This includes supporting satellite data capture and processing, monitor and control of spacecraft and science instruments, prototyping user interface of the Space Station Freedom crew workstations and supporting diagnostic display windows for realtime control systems in ground operations. For these types of applications, TAE Plus is principally used to design and manage the user interface, which is made up of a combination of user entry and data-driven interaction objects. TAE Plus becomes a part of the development life cycle as projects use TAE Plus to prototype the initial user interface design and have this designed user interface evolve into the operational UI.

Outside the NASA community, TAE Plus is being used by an assortment of other government agencies (13%), universities (15%), and private industries (40%). Within the government sector, users range from the National Center for Atmospheric Research, National Oceanographic and Atmospheric Adminstration, U.S. Geological and EROS Data Center, who are developing scientific analysis, image mapping and data distribution systems, to numerous Department of Defense laboratories, who are building command-and-control systems. Universities represented among the TAE community include Cornell, Georgia Tech, MIT, Stanford, University of Maryland and University of Colorado. Applications being developed by University of Colorado include the Operations and Science Instrument Support System(OASIS), which monitors and controls spacecraft and science instruments and a robotics testbed for research into the problems of construction and assembly in space.[10] Private industry has been a large consumer of the TAE technology and a sample of the companies that have received TAE Plus include Apple Computer Inc., Loral Aerospace, Martin Marietta, Computer Sciences Corp., TRW, Lockheed, IBM, Northern Telecom, Mitre Corp., General Dynamics and GTE Government Systems. These companies are using TAE Plus for an assortment of applications, ranging from a front-end for a corporate database to advanced network control center. Northern Telecom, Inc. used TAE Plus to develop a technical assistance service application which enables users to

easily access a variety of applications residing on a network of heterogeneous host computers.[11] Because of the high cost associated with programming and software-development, more and more software development groups are looking for easy-to-use productivity tools, and TAE Plus has become recognized as a viable tool for developing an application's user interface.

NEXT STEPS

The current TAE Plus provides a useful tool within the user interface development environment -- from the initial design phases of a highly interactive prototype to the fully operational application package. However, there are many enhancements and new capabilities that will be added to TAE Plus in future releases.

In the near term, the emphasis will be on enhancements and extensions to the WorkBench. All the requested enhancements are user-driven, based on actual experience using TAE Plus, or requirement-driven based on an application's design. For example, on the enhancements list are extensions to the interaction objects, (e.g., graph data-driven object, form fill-in), support for importing foreign graphics, refinements in the code generation feature, extensions to the connections feature (e.g., graphic representation of the connection mapping, item-to-item connections), and multiple console support.

Future advancements include expanding the scope of TAE Plus to include new tools and technologies. For instance, the introduction of hypermedia technology and the integration of expert system technology to aid in making user interface design decisions are targeted for investigation and prototyping.

CONCLUSION

With the emergence of sophisticated graphic workstations and the subsequent demands for highly interactive systems, the user interface becomes more complex and includes multiple window displays, the use of color, graphical objects and icons, and various selection techniques. Software tools, such as TAE Plus, are providing ways to make user interface developer's tasks easier and improve the overall productivity of the development process. This includes supporting prototyping of different user interface designs, as well as development and management of the operational application's user interface.

TAE Plus is an evolving system, and its development will continue to be guided by user-defined requirements. To date, each phase of TAE Plus's evolution has taken into account advances in windowing systems, human factors research, command language design, standardization efforts and software portability. With TAE Plus's flexibility and functionality, it is providing a useful productivity tool for building and managing graphical user interfaces.

ACKNOWLEDGEMENTS

TAE Plus is a NASA software product being developed by the NASA/Goddard Space Flight Center with prime contract support by Century Computing, Inc. The work is sponsored by the NASA Office of Space Operations.

TAE is a registered trademark of National Aeronautics and Space Administration (NASA). It is distributed through NASA's distribution center, COSMIC, (404) 542-3265. For further information, contact COSMIC and/or the TAE Support Office at GSFC, (301) 286-6034.

REFERENCES

1. Perkins, D.C., Howell, D.R., Szczur, M.R., "The Transportable Applications Executive -- an interactive design-to-production development system," *Digital Image Processing In Remote Sensing*, edited by J-P Muller, Taylor & Francis Publishers, London, 1988.

2. Scheifler, Robert W., Gettys, Jim., "The X Window System," MIT Laboratory for Computer Science, Cambridge, MA, October 1986.

3. Open Software Foundation, Inc., *OSF/Motif™ Programmer's Reference Manual*, Revision 1.1, 1990

4. Cox, Brad J., *Object Oriented Programming, An Evolutionary Approach*, Addison-Wesley Publishing Company, Reading, MA, 1986.

5. Stroustrup, Bjarne, *The C++ Programming Language*, Addison-Wesley Publishing Company, Reading, MA, 1987.

6. Szczur, Martha R., Miller, Philip, "Transportable Applications Environment (TAE) Plus: Experiences in 'Object'ively Modernizing a User Interface Environment," Proceedings of the OOPSLA Conference, September 1988.

7. Linton, Mark A., Vlissides, John M., Calder, Paul R., "Composing User Interfaces with Interviews," *IEEE Computer*, February, 1989.

8. Evans, M.W., Piazza, P., Dolkas, J.B., *Principles of Productive Software Management*, John Wiley and Sons Publishers, 1983.

9. Boehm, Barry, "Improving Software Productivity", *IEEE Computer*, September, 1987, pp. 43-57

10. Klemp, Marjorie, "TAE Plus in a Command and Control Environment", Proceedings of theTAE Eighth Users' Conference, June, 1990

11. Sharma, Alok, et al., "The TAS Workcenter: An Application Created with TAE", Proceedings of the TAE Eighth Users' Conference, June, 1990

N92-22718

INSTRUMENTATION, PERFORMANCE VISUALIZATION AND DEBUGGING TOOLS FOR MULTIPROCESSORS

Jerry C. Yan and Charles E. Fineman
Sterling Federal Systems Inc.
MS 244-4, NASA Ames Research Center, Moffett Field, CA 94035

Philip J. Hontalas
Computational Systems Research Branch
MS 244-4, NASA Ames Research Center, Moffett Field, CA 94035

ABSTRACT

The need for computing power has forced a migration from serial computation on a single processor to parallel processing on multiprocessor architectures. However, without effective means to monitor (and visualize) program execution, debugging and tuning parallel programs becomes intractably difficult as program complexity increases with the number of processors. Research on performance evaluation tools for multiprocessors is being carried out at NASA Ames Research Center. Besides investigating new techniques for instrumenting, monitoring and presenting the state of parallel program execution in a coherent and user-friendly manner, prototypes of software tools are being incorporated into the run-time environments of various hardware testbeds to evaluate their impact on user productivity. Our current tool set, the Ames InstruMentation System (or AIMS), incorporates features from various software systems developed in academia and industry. The execution of FORTRAN programs on the Intel iPSC/860 can be automatically instrumented and monitored. Performance data thus collected can be displayed graphically on workstations supporting X-Windows. We have successfully compared various parallel algorithms for CFD applications in collaboration with scientists from the Numerical Aerodynamic Simulation Systems Division. By performing these comparisons, we show that performance monitors and debuggers such as AIMS are practical and can illuminate the complex dynamics that occur within parallel programs.

1. INTRODUCTION

1.1. Motivation and Background

While parallel processing promises to deliver orders of magnitude speed-up in the near future, the actual speed-up obtained from parallel processing will always depend critically on three factors: i.) how the parallel application is formulated; ii.) the architecture of the multiprocessor and iii.) how well the application is mapped onto the machine. Although research in these areas has produced many interesting results based on simulation and theoretical considerations, their validity must be substantiated by data gathered from actual implementations. Such performance evaluation on multiprocessors presents many technical challenges.

A parallel program has many threads of control. Whether they are expressed as "parallel do-loops" or concurrent processes/objects, the completion time of the entire program depends on the order in which synchronization/communication events occur on different control threads. This "event-ordering" data is difficult to collect, analyze and present in a manner that relates performance with program structure and hardware architecture. Having accurate resource utilization information, for example, can be especially helpful for evaluating the effectiveness of the current program-to-machine mapping — whether there is proper trade-off between communication and concurrency as the computation is distributed over many processors.

In summary, whether a researcher is designing the "next parallel programming paradigm", another "scalable multiprocessor" or investigating resource allocation algorithms for multiprocessors, a facility that enables parallel program execution to be captured and displayed is invaluable. Careful analysis of such infor-

mation can help computer and software architects to detect, and therefore, exploit behavioral variations among/within parallel programs to take advantage of specific hardware characteristics.

1.2. Instrumentation Methodologies

Performance evaluation presumes some form of *instrumentation* — a mechanism whereby the execution of the program can be monitored. A variety of such mechanisms have been proposed to gather different information; these include *event sampling*, *passive event recorders*, and *inserted active event recorders*. A detailed survey of the various instrumentation methodologies for multiprocessors may be found in [1].

An event sampler, whether software or hardware, periodically examines and records the state of the executing software. For example, the UNIX *gprof* [2] has been used to collect statistics about the the distribution of work among the modules and statements of a sequential application. In a sequential environment, an external agent (usually another process in a multiprogramming environment) carries out the *sampling* by periodically interrupting the monitored process to record the value of its program counter. Based on the data collected, the time spent in various parts of a program can be determined. *Event sampling* techniques have been applied successfully on sequential programs for many years now. In a parallel processing environment however, *event sampling* might not be feasible because a *sampling process* can be highly intrusive. Even if the problem of intrusion is overcome through the use of specialized instrumentation hardware, the inter-process event dependencies often found in parallel programs cannot be reconstructed based on statistical data alone.

The use of passive event recorders requires specialized instrumentation hardware for implementation. The word "passive" implies that a monitored system does not do anything *extra* for performance data to be collected. Program state, therefore, must be deduced from low-level data gathered from various devices such as addresses/data placed on buses or values in registers. Even with simple sequential programs, a large amount of data has to be gathered. This implies that instrumentation hardware for parallel systems has to cope with even higher data rates and capacities. Furthermore, hardware monitors tend to be inflexible and vendor specific. The algorithms that relate collected data to program source code must take into account specific compilation strategies and operating system versions. It takes a lot of effort to build a single passive instrumentation system — not to mention building a suite across different software/hardware architectures for research and development.

Inserted active event recorders collect exactly what you want to measure — no more no less. Just like putting print-statements at various points in the program to trace its control-flow, "event records", which indicate event types and their times of occurrence, can be placed at various points of the source code. Program execution can then be easily reconstructed based on these records. The tedious task of instrumenting program source code can be automated, even across different parallel programming languages[1]. Furthermore, this approach is highly portable since the program is instrumented at the source code level. The performance of an instrumented parallel program can be studied on any machine without major modification. Because the event format can be standardized across different machines/languages, only one set of performance analysis tools is required to interpret the data gathered. Although the overhead of this approach is not negligible, it still can be accurately measured, characterized and factored out using various compensation techniques (e.g. [3]).

1.3. Outline of Paper

The goal of this paper is to present some of the techniques and methodologies employed in the instrumentation and performance debugging of applications executing on multiprocessors. To that end, this paper will present our current tool set, the Ames InstruMentation System (AIMS), as an example. Section 2 of the paper describes how AIMS monitors program execution. The *source code instrumentor* automatically inserts active event recorders (i.e. subroutine calls to the *run-time performance monitoring library*) into the source code before compilation. Performance data generated by these event recorders are gathered into a *trace file* from which the *visualization tool-set* reconstructs program execution. Section 3 contains a sample of views obtained

---

[1] For example, PIE [7] uses a source code instrumentor that handles parallel programs written in C, Ada, and FORTRAN.

using AIMS to measure the performance of a parallel version of ARC2D, a computational fluid dynamics application, on the Intel iPSC/860 at NASA Ames Research Center. In this case, AIMS helped the researcher to idenfify execution bottlenecks and room for improvement. Conclusions and directions for future research are discussed in section 4.

2. THE AMES INSTRUMENTATION SYSTEM

2.1. Structural Overview

AIMS is designed to facilitate performance evaluation of parallel applications on multiprocessors by capturing and visualizing execution data. AIMS has three major software components: a *source code instrumentor*, a *run-time performance monitoring library* and a *visualization tool-set*.

The instrumentor inserts active event recorders (i.e. function calls to the monitor library) directly into the application source code with little or no intervention by the user. AIMS provides a graphical interface for the researcher to selectively instrument his/her code. As shown in Figure 1, specific modules and procedure calls can be selected/deselected easily via the click of a mouse. Thus, the programmer is relieved of the tedious work of instrumentation by hand.

The monitor library provides a set of active event recorders to measure and record various aspects of program performance such as message passing overhead, processor synchronization overhead, and processor time spent in user defined areas of the application.

The visualization tool-set processes the execution data gathered and displays them using graphical views. Detailed information showing how the application interacted with the multiprocessor is presented using *animated views*, from which processor state, implementation bottlenecks and load imbalances can easily be observed. Performance statistics of the entire program execution can also be gathered and displayed via *statistical views* to provide insights into the general behavior of the program; these may yield valuable clues regarding where the animated views should be focused.

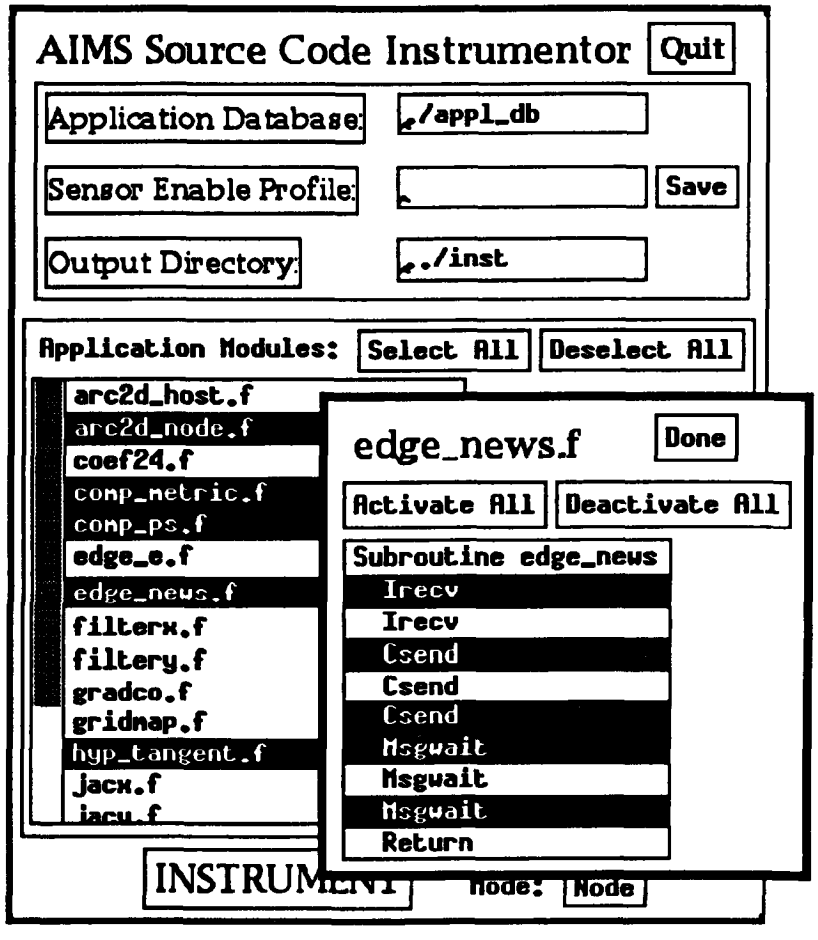


Figure 1. Graphical Interface to AIM's *Source Code Instrumentor*

2.2. Usage Overview

By applying each of the AIMS components sequentially, the performance of various parallel programs on a multiprocessor can be evaluated. As shown in Figure 2, the source code is first instrumented automatically by

AIMS's *instrumentor*. By default, points of interest include message sending, receiving and blocking as well as procedure entries and exits. The user may specify the procedures and code blocks to be monitored, as well as other instrumentation parameters, via a configuration file. Besides adding code at various points in the source code to generate event records, some system calls are replaced by monitor library calls when timing measurements have to be made within such calls[2]. After the source code is instrumented, it is compiled and linked with the run-time performance monitoring library.

The instrumented program is then loaded and run on the multiprocessor. Performance data is gathered during program execution and stored to local memory buffers. Periodically, these buffers fill up and the data is written out to a *trace file* on the file system. Event records generated include:

- *procedure events* — provide performance data on user selected subroutines;
- *blocking events* — indicate waiting time spent on synchronization;
- *message events* — records message transmission time, message size, destination and type; and
- *statistical event records* — summarizes cumulative performance statistics at specified points of program execution.

Finally, the *trace file*, which contains the event records for a monitored program execution, is collected and transferred to a graphic work station to be processed and displayed in various formats. The *visualization toolset* reads the performance data from the trace file and interprets that information on a variety of X-window based displays. With the aid of an example, we will illustrate how different displays can capture various aspects of system performance in section 3.

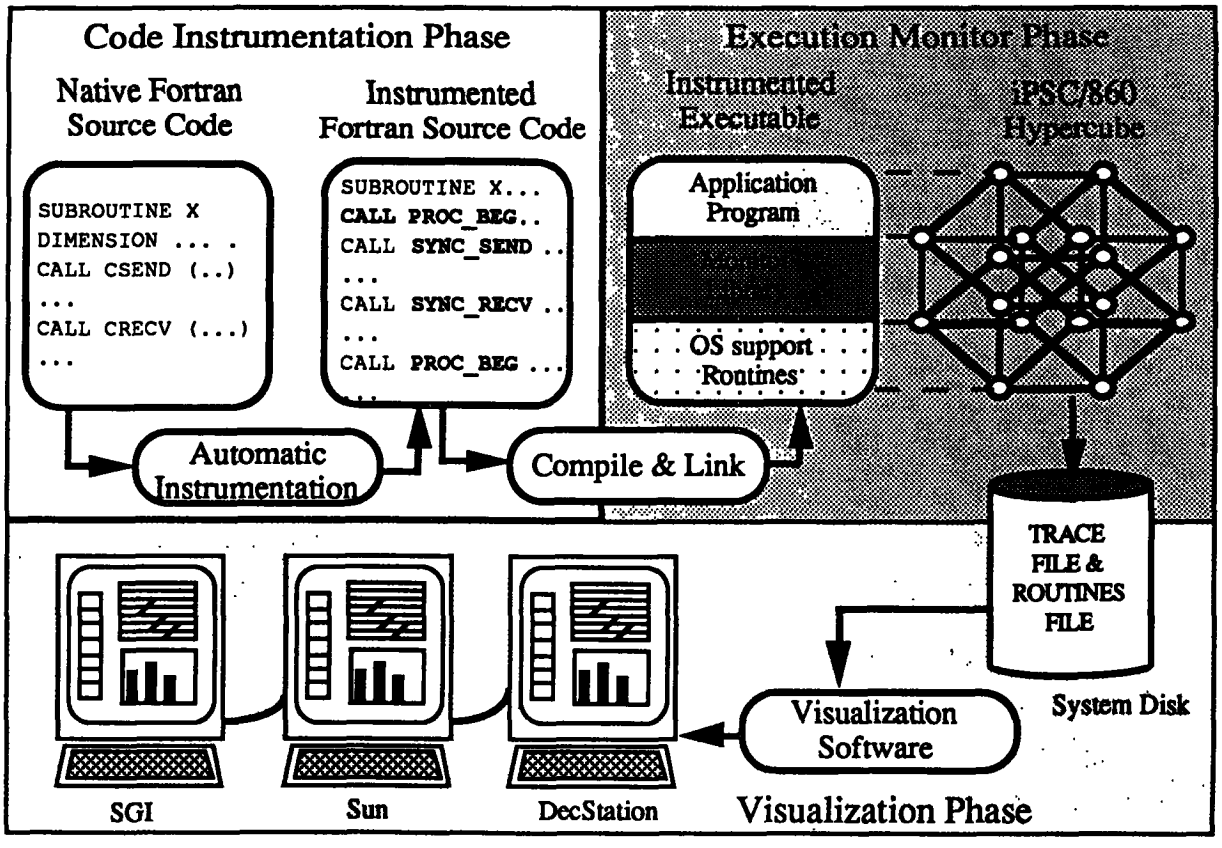


Figure 2. Using AIMS to Collect Performance Data

[2] For example, SYNC_SEND and SYNC_RECV replaces CSEND and CRECV on the Intel iPSC/860 while at the same time, providing timing data about this message transaction.

3. VISUALIZING PARALLEL PROGRAM EXECUTION

3.1. The Example Application

A grand challenge of NASA's High Performance Computing and Communications Program [4] involves the development of parallel Computational Fluid Dynamics (CFD) programs. CFD involves the numerical solution of a system of nonlinear partial differential (Navier-Stokes) equations — these represent the laws of conservation of mass, momentum and energy applied to a fluid medium in motion. One such FORTRAN program, ARC2D [5], which applies an implicit solution algorithm to a problem with two spatial dimensions, has been parallelized for the Intel iPSC/860 Hypercube (an MIMD multiprocessor) at NASA Ames Research Center.

3.2. A Few Examples of Animated Views

The AIMS *visualization toolset* was developed after a careful evaluation of the views provided by the *ParaGraph* [6] visualization toolset and PIE [7]. We selected those that we found useful for our applications and incorporated them into AIMS. In this paper, we only describe those views that are not provided by ParaGraph. The *OverVIEW Diagram* shown in Figure 3 animates program execution by scrolling from right to left. When a processing node (say #15) is busy, a colored bar is drawn (next to the label "15"). The bar is colored according to the subroutine currently executing. White space indicates that the processing node is idle, probably waiting for the arrival of a message. When a message is passed (say from #15 to #14), a (blue) line is drawn from the point (on the sender's time line) when the message was sent to the point (on the receiver's time line) when the message was removed from the queue. The *Aggregate Processor Utilization Chart* plots processor utilization as a function of time. The height of the curve denotes the number of processors currently busy. As shown in Figure 4, it is also color-coded according to subroutine name.

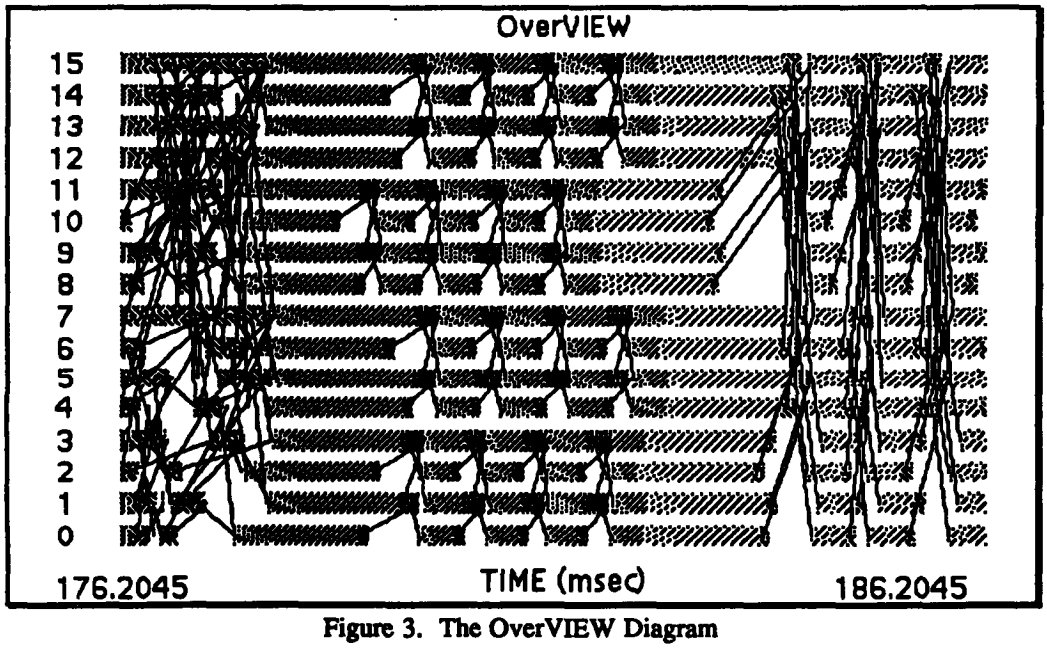


Figure 3. The OverVIEW Diagram

Besides providing views focused on the parallel program's flow of control, AIMS also provide views that display the state of each processor at particular points in time. The Grid view shown in Figure 5 is such an example. Each box of the Grid view displays the current state, subroutine being executed, message queue size and overall utilization for each processor. In addition, this view permits the developer to map the physical processors of the iPSC/860 onto a two dimension mesh. Many parallel applications (such as ARC2D) can be decomposed to topologies which may not conform exactly to the iPSC/860's hypercube.

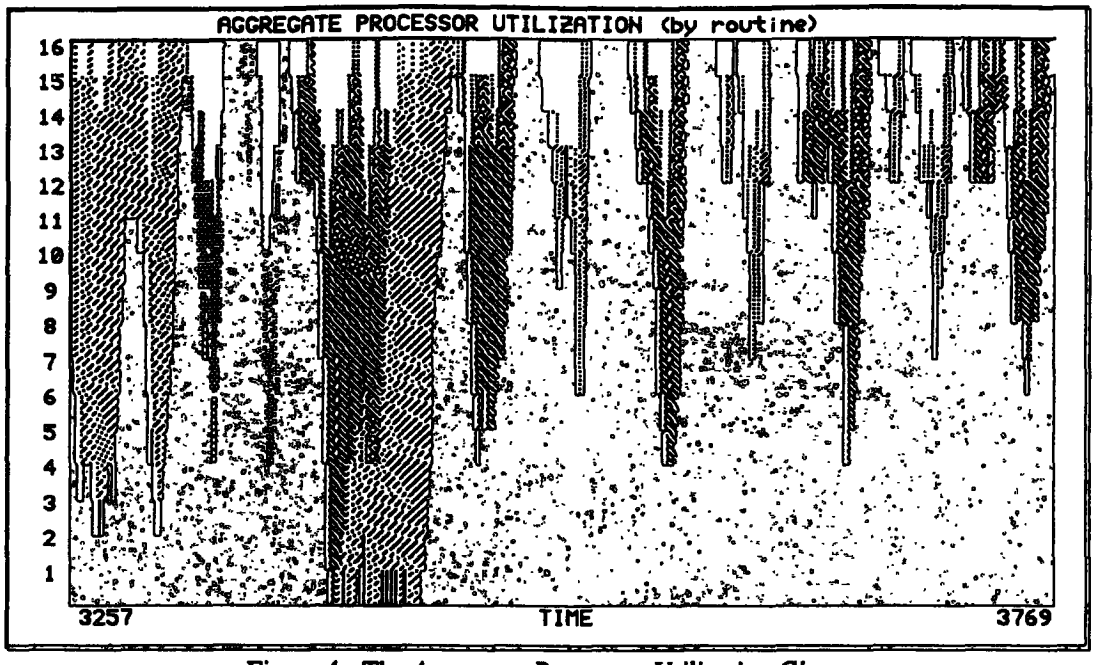


Figure 4. The *Aggregate Processor Utilization Chart*

The NCPU view summarizes the performance characteristics for the entire execution. As shown in Figure 6, a histogram plots the normalized[3] CPU usage of various subroutine. For example, `ypldge` spends most of its time executing when 12 processors are busy.

Based on these animation and statistical views, the programmer can identify the subroutines and message transactions associated with periods of idleness in his/her program. This, in turn, provides valuable insights about the parallelization strategy chosen and helps the programmer to reformulate the application if necessary.

Besides providing graphical data for performance tuning, AIMS also provide an important feature known as *source code click-back*. A mouse click in the OverVIEW will bring up

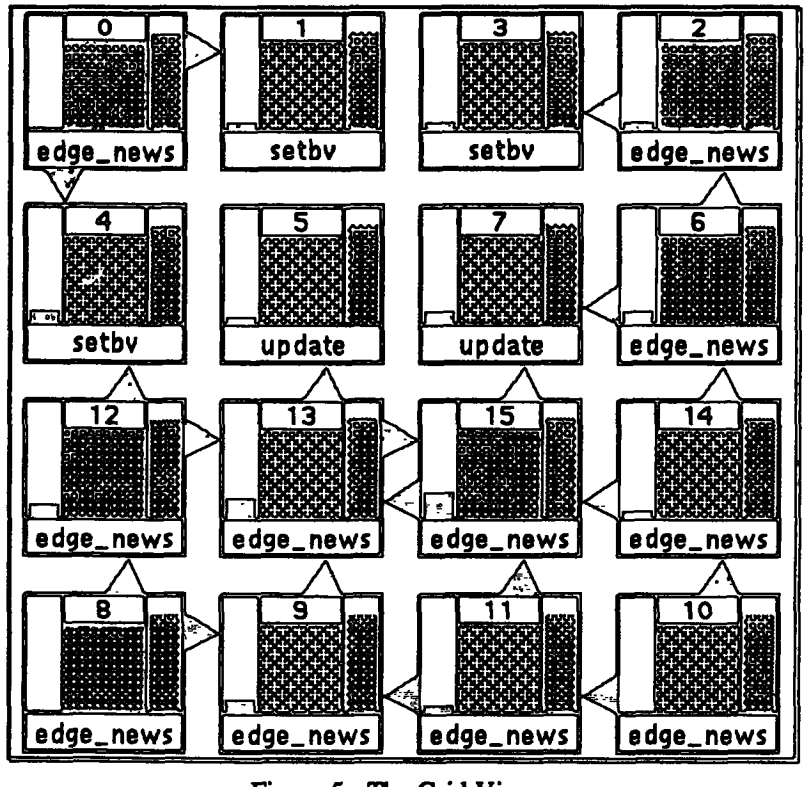


Figure 5. The Grid View

[3] The normalized CPU usage of a subroutine is the total amount of CPU time it used divided by k where k processors were active simultaneously.

a text window depending on the location of the cursor in the view. If the cursor was pointing to a message line, the text file containing the `send` command will be opened and the corresponding program line will be highlighted (as shown in Figure 7). If the cursor is pointing to an idle period of the processor and this idling was caused by the late arrival of a message, the exact `msgwait` call responsible will also be identified. Finally, if the mouse is clicked over a color bar, the code for that subroutine will be retrieved.

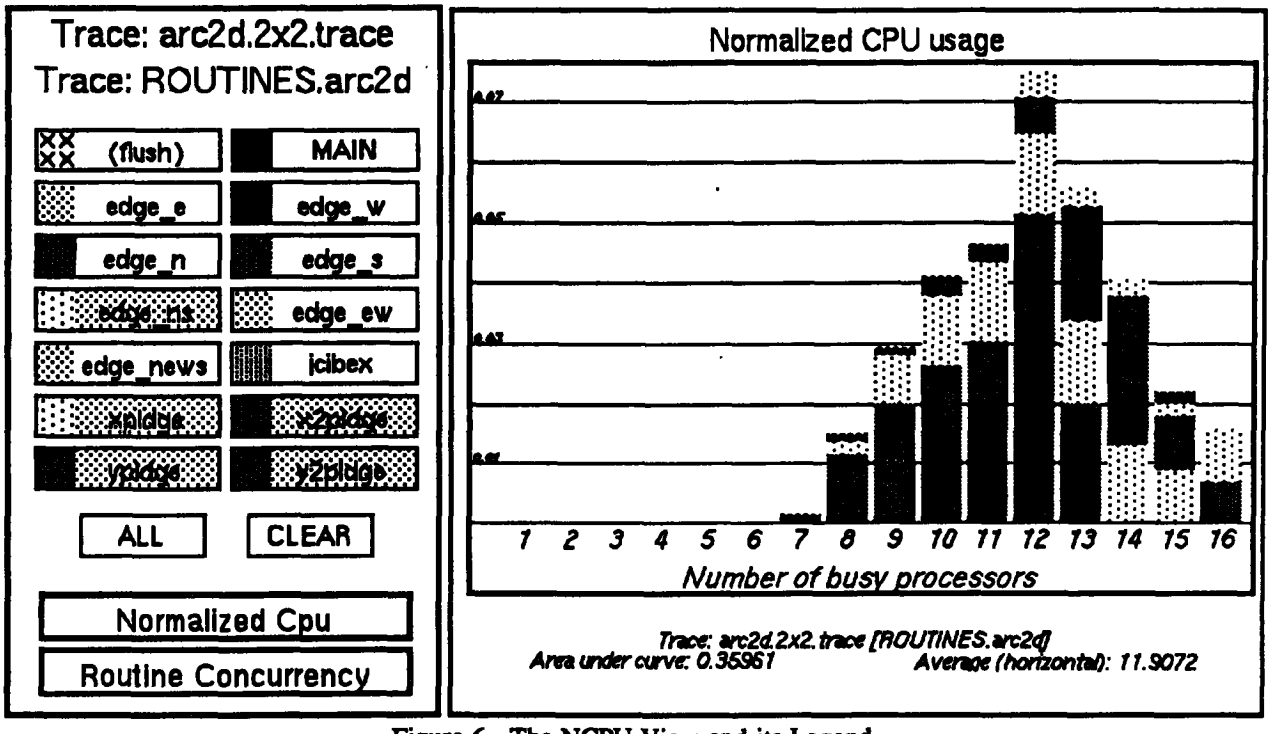


Figure 6. The NCPU View and its Legend

4. CONCLUSIONS AND FUTURE RESEARCH

In summary, the Ames InstruMentation System provides a suite of software tools to facilitate the tuning and debugging of parallel applications. FORTRAN source code is instrumented automatically. Performance data gathered from the execution of instrumented code can be displayed on a variety of workstations. These displays may provide researchers a means for observing the behavior of their programs as well as tracing the sequence of operations via "source code click-back". Thus the performance and correctness of parallel algorithms on hypercubes may be evaluated easily.

Although we have shown that AIMS can be a powerful tool for the development of parallel applications, it is not without pitfalls. One major obstacle to be overcome is data size. Programs running on parallel processors tend to produce an enormous amount of performance data using the techniques described here. Furthermore, data written to disk asynchronously from each processor must be sorted by execution time before it can be read by the visualization toolset. If the data set is particularly large, the overhead of processing this data could render the tools described here impractical. Our current research efforts are addressing the data size and sorting problem from several directions. These solutions include:

- refining the instrumentor to be more selective about which portions of the program to monitor. In future versions of the AIMS, the researcher will be able to enable and disable monitoring according to time and processor parameters. This approach has the potential of greatly reducing the data size.
- integrating a merge sort of the raw data from the multiprocessor at the time of visualization. This technique will eliminate the time consuming pre-sorting process by performing a merge sort of the raw data streams coming from each processor.

- developing "course grain" monitoring tools to compliment the fine grain monitoring capabilities of AIMS. The development of these tools will permit the developer to get a coarse grain view of an application's performance behavior for sampled time periods. Such an approach should have lower overhead in terms of data collection. Based on these coarse grain views, the researcher may identify problem spots quickly which can then be examined more closely be the fine grain performance monitoring facilities of AIMS.

Finally, all performance monitoring systems must deal (to one extent or another) with the problem of perturbance. Instrumentation overhead may re-order events in different control threads of a parallel program and, therefore, obscure the actual data collected. Future versions of AIMS will produce statistics that help determine the level of perturbance within the monitoring process and compensate the performance displays accordingly.

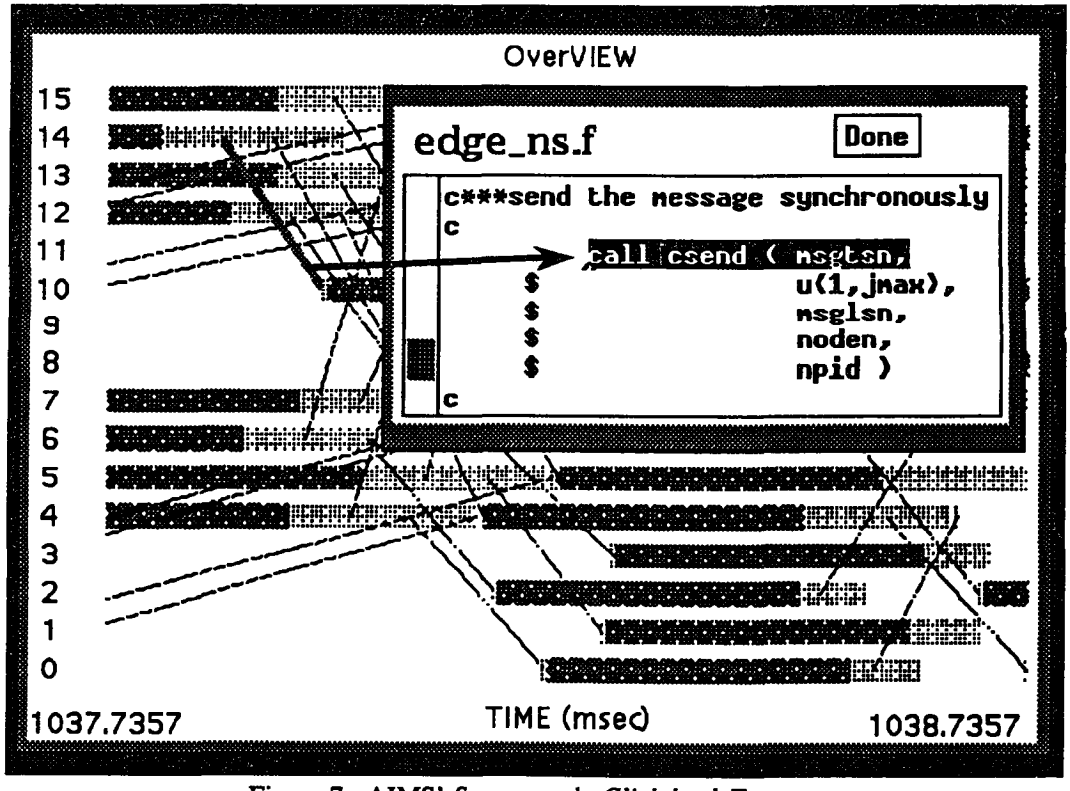


Figure 7. AIMS' *Source-code Click-back* Feature

ACKNOWLEDGEMENTS

AIMS was put together after careful evaluation of a few software prototypes from the research community as well as published ideas on performance visualization. We would like to acknowledge their help/support in letting us adopt, adapt and augment their research prototypes for the parallel processing environment here at NASA Ames Research Center. We also want to thank two summer students, Chris Hanson from Santa Clara University and Sheralyn Listgarter from Stanford University who spent many hours implementing (and watching) scrolling and flickering windows on our workstation screens.

AIMS' instrumentor is implemented on top of a parser developed for PIE [7] which we obtained through the CMU Affiliates' Program. AIMS's monitor library adopted many of the interface conventions established by PICL — a Portable Instrumented Communication Library [8] developed at Oak Ridge National Laboratory. AIMS's visualization toolset incorporated display concepts use by AXE [9] from NASA Ames Research Center, ParaGraph [6] from Oak Ridge National Laboratory and *Quartz* [10] from University of Washington.

REFERENCES

[1] M. H. Reilly, *A Performance Monitor for Parallel Programs*. Academic Press Inc. 1990.

[2] S. L. Graham, P. B. Kessler, M. K. McKusick. "gprof: a Call Graph Execution Profiler". In *Proceedings of SIGPLAN '82 Symposium on Compiler Construction*, June 1982.

[3] T. Lehr. *Compensating for Perturbation by Software Performance Monitors in Asynchronous Computations*. Ph.D. Dissertation, Department of Electrical Engineering, Carnegie Mellon University. 1990.

[4] *Grand Challenges: High Performance Computing and Communications*. A Report by the Committee on Physical, Mathematical, and Engineering Sciences, Federal Coordinating Council for Science, Engineering, and Technology, Office of Science and Technology, Executive Office of the President. 1991.

[5] D. Bailey, J. Barton, T. Lasinski, and H. Simon Ed., "The NAS Parallel Benchmarks" Report RNR-91-002, NAS Systems Division, NASA Ames Research Center. January 91.

[6] M. Heath and J. Ethridge. "Visualizing the Performance of Parallel Programs". *IEEE Software*, Vol. 8, No. 5, Sept. 1991, pp. 29-39.

[7] Z. Segall, L. Rodolph, "PIE — A Programming and Instrumentation Environment for Parallel Processing," *IEEE Software*, Vol. 2, No. 6, Nov. 1985, pp. 22-37.

[8] G. A. Geist, M. T. Heath, B. W. Peyton, P. H. Worley "PICL — A Portable Instrumented Communication Library" Tech Report ORNL/TM-11130, Oak Ridge National Laboratory. May 1990.

[9] J. C. Yan. "Axe — An Experimentation Environment for Concurrent Systems". *IEEE Software*, page 25, May 1990.

[10] T. E. Anderson and E. D. Lazowska. "Quartz: A Tool for Tuning Parallel Program Performance". In *Proceedings of SIGMETRICS '90 Conference on Measurement and Modeling of Computer Systems*, May 1990, pp. 115-125.

N92-22719

OPERATIONS AUTOMATION USING TEMPORAL DEPENDENCY NETWORKS

Lynne P. Cooper
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

ABSTRACT

Precalibration activities for the Deep Space Network are time- and work force-intensive. Significant gains in availability and efficiency could be realized by intelligently incorporating automation techniques. This paper presents an approach to automation based on the use of Temporal Dependency Networks (TDNs). A TDN represents an activity by breaking it down into its component pieces and formalizing the precedence and other constraints associated with lower-level activities. This paper describes the representations used to implement a TDN and the underlying system architecture needed to support its use. The commercial applications of this technique are numerous. It has potential for application in any system which requires real-time, system-level control and accurate monitoring of health, status, and configuration in an asynchronous environment.

INTRODUCTION

During precalibration (precal) of the Deep Space Network (DSN), operators configure the required subsystems (e.g. antennas, receivers, transmitters), download support data (e.g. standards and limits, pointing coordinates), and perform tests and calibration procedures. To do this, DSN operators send hundreds of directives and monitor the 1000+ response messages generated by the subsystems. Efforts have been made to automate portions of precals through software macros: creating a list of directives and then executing them through the macro rather than individually. This approach has fallen short due to two major limitations: (1) the inability to explicitly specify all possible contingencies, and (2) the lack of visibility into system status should the macro fail in-process.

Our approach is to represent the procedures as a temporal dependency network (TDN) where high-level procedures are represented as logical nodes of the network. The nodes consist of the directives needed to accomplish the task, temporal constraints, pre- and post- conditions, and local recovery information should the node "fail". The network specifies precedence relationships between nodes, any potential parallelism, and rules for recovering from global faults. The network is executed using a blackboard architecture which performs special monitoring functions so that the operator is always aware of the status of the equipment and the executing procedures.

The commercial applications of this technique are numerous. It has potential for application in any system which requires real-time, system-level control and accurate monitoring of health, status, and configuration in an asynchronous environment. This paper will first discuss the characteristics of the problem domain which led to the use of the TDN. It will then present the knowledge representation techniques used for the Temporal Dependency Network and the Domain Model. It will then discuss the system architecture of the Operator Assistant which uses the TDN to support semi-autonomous operations. Finally, it will discuss some of the lessons learned during the early prototyping stages and plans for demonstration of the technology in an operational environment.

DOMAIN ANALYSIS

DSN Link Monitor & Control

The operators at the DSN antenna complexes are responsible for setting up all of the equipment necessary to provide a communications path ("Link") between an individual spacecraft and its mission operations center[1]. In performing this function, the DSN essentially acts as a "bent pipe" -- passing through the information in a way which is philosophically similar to the way the Postal Service handles our mail. For the DSN, the goal is to deliver the data from the spacecraft (or commands to the spacecraft) as reliably as possible and for as many spacecraft as possible.

Currently, the DSN operators spend a significant amount of time performing the setup (precal) functions necessary to support a link. The Link Monitor & Control (LMC) operators initialize, configure, test, calibrate, and otherwise prepare the equipment by manually entering directives to each of the subsystems through an LMC console[2]. In order to effectively perform their jobs, the LMC operators must learn hundreds of different directives, interpret more than a thousand different messages, and evaluate 100 different information displays. This is more difficult than expected because each subsystem has its own unique "language", all directives must be entered in a cumbersome syntax through a keyboard, and the procedures themselves differ from antenna to antenna, and mission to mission. Furthermore, these procedures are not fully integrated. The spacecraft user guides provide a linear overview of the procedures necessary for their projects, but require that the LMC operators integrate other procedures available only from the subsystem handbooks. In addition, numerous changes to the procedures are received from engineering and science personnel and must be incorporated. To further complicate matters, the values needed to parameterize the directives for the specific track are also spread throughout volumes of documentation and numerous printouts of schedules and other support data.

A significant percentage of DSN operations time is spent performing precals. Reducing this operations overhead, and thereby increasing the availability of the DSN, is a major goal of the applied research presented in this paper. The two concepts which guide the effort are Positive Control and Situation Awareness.

Positive Control

In this context, positive control refers to the desire that all control actions (i.e. directives) have explicit feedback as to their effects . Under the existing LMC system, a limited set of messages report the status of directives. Unfortunately, these messages are lumped together with all of the other messages (monitor information, events, alarms, warnings, status announcements). It is up to the operator to "catch" the response as it scrolls by on a very busy text display.

Situation Awareness

There are two facets to situation awareness: (1) knowing the configuration, health, status, and readiness of the link equipment; and (2) knowing the status of the execution of the required procedures. The first of these is an information fusion problem requiring that the pertinent information be available and presentable to the operator. The second requires that the operator have an explicit high-level plan, and that the operator knows where s/he is at all times in the execution of that plan. While there is limited explicit feedback associated with the individual directives, the existing system suffers from the lack of explicit information on the "side effects" of the directives. For example, a directive which turns on the receiver will elicit a response saying that receipt of the directive has been acknowledged by the receiver. However, the operator must use other means to determine if the directive had the required effect, i.e., that the receiver did indeed turn on.

---

[1] While the equipment which makes up the path is referred to as the "Link", the setup process is referred to as precalibration, and the actual beginning-to-end setup, communications, and tear-down is referred to as a "Track" or "Pass."

[2] There are some subsystems which cannot be remotely controlled or monitored. For those subsystems, the operator verbally requests a technician in the actual equipment room to do what is necessary to configure, etc. the equipment.

Problem Summary

Successful automation requires that appropriate thought be given to the point at which the human operator is once again included in the system. Since DSN LMC is a complex, real-world environment, complete with ambiguity, incomplete sensor feedback, and creative combinations of circumstances which make it virtually impossible to specify all possible contingencies, our system is semi-autonomous, requiring a human operator in order to function fully -- but capable of reducing the amount of menial tasks require of those operators.

The requirements (goals) which drive our approach are:

1. Reduce manual input required of the operators (keyboard entries).
2. Provide a high-level representation of the task to be performed by integrating the necessary procedures.
3. Provide a means of easily accessing and integrating support data.
4. Parallelize the execution of the procedures.
5. Provide configuration and procedure situational awareness.
6. Support positive control (to the extent possible in the existing DSN architecture)
7. Support both autonomous and operator-in-the-loop recovery.
8. Provide fault detection, diagnosis, correction, and replanning.

KNOWLEDGE REPRESENTATION

Procedure Representation

The procedure necessary to accomplish a specific pass is represented as a Temporal Dependency Network (TDN). This is essentially a Petri Net which represents precedence relationships between procedures and which overlays time constraints. Each block in the network is a frame-type object which consists of its name, preconditions for its execution, the directives to be executed as part of the block, functions which identify how to fill in the needed parameters, time constraints (e.g. acquire the spacecraft at time, t), and the postconditions which exist after the block has been completed.

Each block represents a set of directives to be executed sequentially. The preconditions which must be met before the block can be executed are valid for the entire block. Similarly, the post conditions are valid only after completion of the entire block, although only one directive may be responsible for creating that condition. These decisions were made to simplify the execution of the TDN. In cases where existing procedures violated these requirements, the logical operations were broken down into component pieces to ensure that the pre- and post-condition representations were consistent with the design decisions described above. For example, if the existing, accepted way to perform a function requires two directives to be sent to the same subsystem sequentially (Turn on antenna hydraulics, Move antenna), but the second must wait for the first to be processed and the action invoked by it to be accomplished completely before beginning (It takes several minutes before the hydraulic system is primed and able to support moving the antenna); then, these two inter-related actions were separated into two separate blocks with explicit precedence relationships and additional constraints related to the system state.

The TDN is a network of the blocks as identified above. It presents the high-level overview of the task as an integrated whole, rather than as disassociated sets of directives. The network explicitly describes how operations of the individual subsystems must be integrated. The TDN also specifies what functions are mandatory for the given track, which are desired, and which should be done only under special circumstances. Figure 1 shows the top-level TDN for Very Long Baseline Interferometry (VLBI). Raw data from a VLBI pass is used to determine accurate positioning information for a spacecraft. Several procedures included in the network are desired, but optional. For example, the scientists analyzing the data want the operators to perform a coherency test prior to beginning the actual data collection. This function is not essential; even if it is not performed, the scientists will still get usable data. However, if the coherency test is performed, they can use the information from the test to better process the data. The TDN differentiates between the precedence relationship between this block and others in the network. Under nominal circumstances, the coherency test block would be executed. If, however, circumstances require a quicker completion of the network as a whole, that block can be bypassed with an acceptable impact on the schedule.

Domain Model

The domain model is the equipment analogue to the procedures represented in the TDN. The domain model contains a description of each piece of equipment which is used in the link. This description includes its configuration (e.g. switch settings), and performance parameters (e.g. Receiver In Lock, Signal to Noise Ratio). The domain model representation is tightly coupled to the TDN because the status of the link is a function of the operations being performed through the TDN. Therefore, the domain model can be thought of as a time series of snapshots which correspond to the expected state of the system as it responds to actions invoked by the different directives.

The domain model also represents the link equipment at higher levels of abstraction than simple configuration and performance parameters alone would allow. For example, the distinction between being just "assigned" to the link vs. "operational" in the link is made. (If equipment is *assigned*, it will accept directives from the LMC operator. When it is operational, it is actively performing its function within the link.)

OPERATOR ASSISTANT ARCHITECTURE

The architecture developed to use the TDN representation is shown in Figure 2. The architecture is interrupt-driven and interprocess communications are in the form of events and messages. Data is passed through a combination of shared memory and datagrams. The primary modules are the Execution Manager, the Response Manager, the Status Manager, the DSN Communications Interface, the Message and Event Router, the Diagnostic Module, and the User Interface. Each module (except for the Comm I/F) contains domain knowledge which is used to guide the decision making processes of the module. The role of each of these will be discussed in the following paragraphs.

Execution Manager

The Execution Manager is responsible for executing the TDN: determining which blocks to activate, setting up queues to handle the directives from the individual blocks, assessing whether preconditions have been met, and responding to any operator overrides or recovery replanning. The EM is an interrupt-driven process. Its primary function it to keep things rolling -- to ensure that the parallelism inherent in the operations procedures is exploited as fully as possible, thereby reducing the amount of time necessary to perform precals.

DSN Communications Interface

As its name implies, the Comm I/F module is responsible for interfacing the Operator Assistant to the operational DSN. Because the communications models used by the DSN are significantly different (philosophically, electronically, logically, ...) from those used internal to the Operator Assistant, this is an extremely important module. It is responsible for physically connecting to the DSN Local Area Network (LAN), accessing information meant for the actual LMC, stripping away the DSN communications protocols and reformatting the data for use by the Operator Assistant. Anyone who has built an extension to an existing system, especially an extension which uses a different design philosophy than the existing system, knows how difficult bridging the gap can be.

Message and Event Router (MER)

The MER takes the output from the Comm I/F and routes it to the appropriate modules within the Operator Assistant. There are several types of messages that the Operator Assistant needs to receive. The MER is responsible for interpreting the message type and parsing the message to extract the important information (and therefore minimize the amount of intermodule traffic). In general, the MER will take the message content and place it in shared memory, then send messages to the appropriate module(s) instructing them to view the message. This approach is currently under evaluation.

Response Manager (RM)

The Response Manager is responsible for matching the directive responses received from the subsystems with the directives sent by the Execution Manager. While most directives have single message responses, in some cases, the subsystem will respond first with an Acknowledgement or Processing message, followed by a Completed

message. In cases of multiple responses, the RM is tasked with ensuring that all responses are attributed to the appropriate directive. The RM is also responsible for detecting timeouts. A response to each directive is expected within a finite length of time (on the order of 5 seconds). If a directive has not been responded to within that limit, the RM must detect the timeout and send an event message to the Execution Manager. In many cases, these timeouts are false alarms, either the LAN lost the message carrying the response, or the time limit was set to an arbitrarily short time. The RM will incorporate knowledge on how to interpret the timeouts.

Status Manager (SM)

The Status Manager is tasked with keeping an up-to-date model of the Link. The SM has a time-tagged model of the expected status of the domain and also maintains an up-to-date model of the actual state of the link. The SM compares the two, notes discrepancies, and takes appropriate action (ignore, send to diagnostic module, ...). The status manager has to accommodate the latency effects inherent in the system. The delay times between the request for action (the directives) and their implicit and explicit effects on the Link are affected by a number of factors such as LAN utilization, subsystem health, physical conditions, and availability of personnel to manually perform a subsystem function. The SM also employs a countdown clock driven by a timed event stack, which provides a means with which to evaluate the overall progress towards completing the precalibration and to determine the probability of success and/or need for corrective action. In general, DSN operations have hard physical constraints such as a limited time window during which a particular antenna can see a spacecraft. The precalibration activities must be complete by the time the spacecraft comes into view, or valuable time may be lost. Since these specific times are known, and the TDN representation allows for estimates of how long activities take, the due times can be propagated backwards to provide a means of determining when an activity is behind schedule.

Diagnostic Module (DM)

The Diagnostic Module incorporates three tightly-coupled functions: (1) Fault Detection, (2) Diagnosis, and (3) Recovery. The other processes initiate the fault detection and diagnosis capability within the DM. The DM pulls information from the other processes as well as the Domain Model representations to accomplish these functions. It then influences the other functions through its recovery module, which contains both correction and replanning functions. Correction functions implement recognized contingency plans. Corrective actions suspend the TDN (or a part of it), jump to an external (to the TDN) procedure, perform those functions, then resume execution of the TDN. To the TDN, correction functions appear to be simple delays.

Replanning is necessary when an event occurs which invalidates a previously accomplished section of the TDN, or requires that TDN execution follow a path different from the nominal one. For example, if a piece of equipment in the link fails, but can be recovered in its pre-failure state by a simple warm start , corrective logic can be invoked. If, however, that piece of equipment fails completely and must be replaced by a brand new piece of equipment, all of the configuration and calibration activities which had taken place using the old equipment must be redone on the new equipment. This requires replanning logic. Assume further that this equipment is optional within the link and the time window constraints for acquiring the spacecraft will be violated before the new equipment is ready. Under these circumstances, the replanning mechanism would be invoked to execute an alternate path of the TDN.

User Interface (UIF)

The final, and probably most critical part of the Operator Assistant, is the User Interface. The UIF will provide the LMC operators with visibility into the automated functions as well as the ability to interrupt, override, or influence those functions, and a mechanism for interacting with all the other elements through graphical user interfaces which support a variety of presentation and input modes. One of the primary functions of the UIF is to make it easier for the operators to do their jobs and to ensure that their ability to function is not unnecessarily or arbitrarily limited by the automation meant to help them. In designing automated systems, there is an almost overwhelming temptation to "fix" individual problems, rather than to view the system as a whole. A good user interface ensures that the human element of the system is as fully integrated as the electronic and mechanical elements.

STATUS AND LESSONS LEARNED

Status

The Operator Assistant has gone through one prototyping phase and is currently being redesigned to address problems which surfaced in the earlier version. A new prototype, built to the architecture presented in this paper is in development and will be tested using DSN compatibility test facilities during the early part of 1992 and demonstrated at an operational site later in 1992. The TDN for VLBI has been reviewed and revised to accommodate additional information. The databases for the directives and their responses are complete and work is now beginning on building the domain model. Currently, the VLBI TDN is the only one in existence. During late 1992, we will conduct an extensibility analysis of the Operator Assistant approach. This analysis will address two problems: (1) how to extend the Operator Assistant to support other types of activities; and (2) how to extend the Operator Assistant to support multiple activities at the same time by one operator. In FY93, following its demonstration within the operational DSN, the Operator Assistant will be moved to the DSN experimental antenna site where it will undergo sustained testing in an operational environment.

Lessons Learned

The past 1-1/2 years of work on the Operator Assistant have yielded some significant results. The first prototype served as a proof of concept -- and also provided some insights which will make the second prototype much more effective. The following are some of the lessons learned highlights from the initial effort.

1. For our domain, a polling-type architecture was inappropriate. During the first prototype, we attempted to make use of a blackboard architecture used to support a robotic vehicle. For that domain, a polling architecture worked well because each of the sensors was continually reporting data, therefore, there was always a data value to be read at each sensor. The Operator Assistant domain, however, was much more asynchronous in nature. Processing time was being wasted polling functions which rarely had anything to report. Also, execution of the TDN was intended to be dynamic so that as preconditions were satisfied, blocks would execute. Processes were spawned and killed routinely and it was difficult incorporating the dynamics of our system into a polling architecture.

2. Integration into an existing system always has "Gotcha's". Although we began our effort with a proven communications interface already in existence and a healthy respect for the difficulties associated with connecting to the DSN -- it was even more difficult than expected -- and for "gotcha" types of reasons: The test area configuration and LAN changed; equipment was down for repair; formal Test Plans were needed afterall; the facilities were not available during normal working hours; a 6-month gap existed between actual operations of the type we wanted to automate; there were serious spacecraft problems, etc.

3. TDNs are useful as a review of operational procedures. During knowledge engineering sessions, the use of a mechanism which explicitly showed the dependencies and parallelism in operations procedures actually uncovered some errors in existing procedures. In some instances, problems surfaced because the subsystem engineers were using a linear method (a listing of directive sequences, complete with GO TO's") to represent a parallel process. Other times, the problems arose because inter-subsystem constraints were not appropriately addressed.

4. TDNs are useful for integrating different operational perspectives. Operational "magic" happens when the efforts of four different groups come together to accomplish a specific purpose: (1) the scientists that use the data; (2) the engineers who design the equipment; (3) the operators who make it all work; an (4) the technicians who fix it when it breaks. In developing the TDN, we met with people from each group -- and each individual had a unique view which enabled the TDN to have a much richer and ultimately more accurate representation.

5. Even in automated systems -- the human element must be integrated as part of the design. In laboratory test cases, it is possible to descope a domain to the point where what remainst can be fully addressed by a system you're building. That luxury doesn't exist in real-world applications. There are always surprises and ambiguities that will crash your system. Building systems that are robust enough to not die -- or at least to die gracefully -- is difficult. But it can be made easier if the human element is integrated as part of the design. For example, the first Operator Assistant prototype ran into network communications problems during a demonstration which we were powerless to overcome due to the fact that there was no interrupt capability available for the operator in that version (high on our list of requirements for the current version!).

6. Make allowances for "system slack" in what you design. In the original prototype, we planned on using the Response Manager to match up both the explicit responses for the directives and the implicit system responses. This approach would have seriously clogged the execution of the TDN because of the latency effects resident in the system - and the rather broad definition of what is nominal performance. Rather than attempting to resolve the ambiguity in absolute terms, we have since decided that we can instead relax the system by using two functions, the Response Manager and the Status Manager.

7. The work put into developing automation is a good foundation for a training system. And vice versa. An independent effort evaluating intelligent training techniques in the same domain as the Operator Assistant began several months ago. By approaching the problem from two different angles, we have set up a very rewarding exchange with the training team. Their foundation work, particularly in the area of the Domain Model is especially applicable.

8. The Absolute Truths.

a. Know and respect your user(s). Remember that they're the ones who will ultimately determine if your system is successful -- and they can help you avoid mistakes.

b. Don't automate the fun stuff. Automation is much more successful if you attack the mundane aspects of the job.

c. Be realistic in your expectations. The operations folks have heard it all (and then some) before. Don't make promises you can't keep. Build a solid foundation and then add to it.

d. Keep in touch with your users. Get their feedback as often as reasonable,

e. Don't bother your users unnecessarily. Remember that they have jobs to do. Do your homework so that you make the best use of the time they can give you.

SUMMARY

The approach to automation described in this paper centers around using Temporal Dependency Networks to represent the procedures to be automated. The commercial applications of this technique are numerous. It has potential for application in any system which requires real-time, system-level control and accurate monitoring of health, status, and configuration in an asynchronous environment.

ACKNOWLEDGEMENTS

The Operator Assistant team has benefited from the contributions of numerous operations, engineering, and scientific personnel, as well as from development team members Elmain Martinez, Juan Urista, Joe Jupin, Rajiv Desai, Lorrine Lee, and Randy Hill.

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

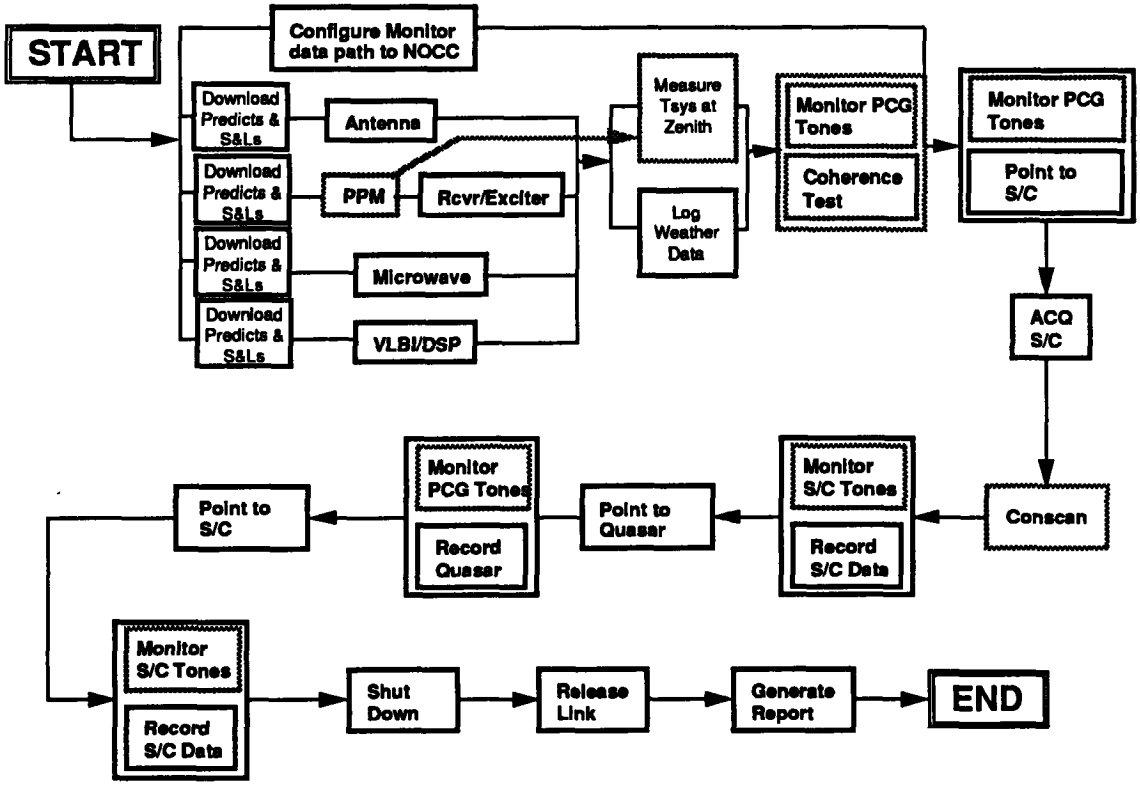


Figure 1. VLBI Temporal Dependency Network

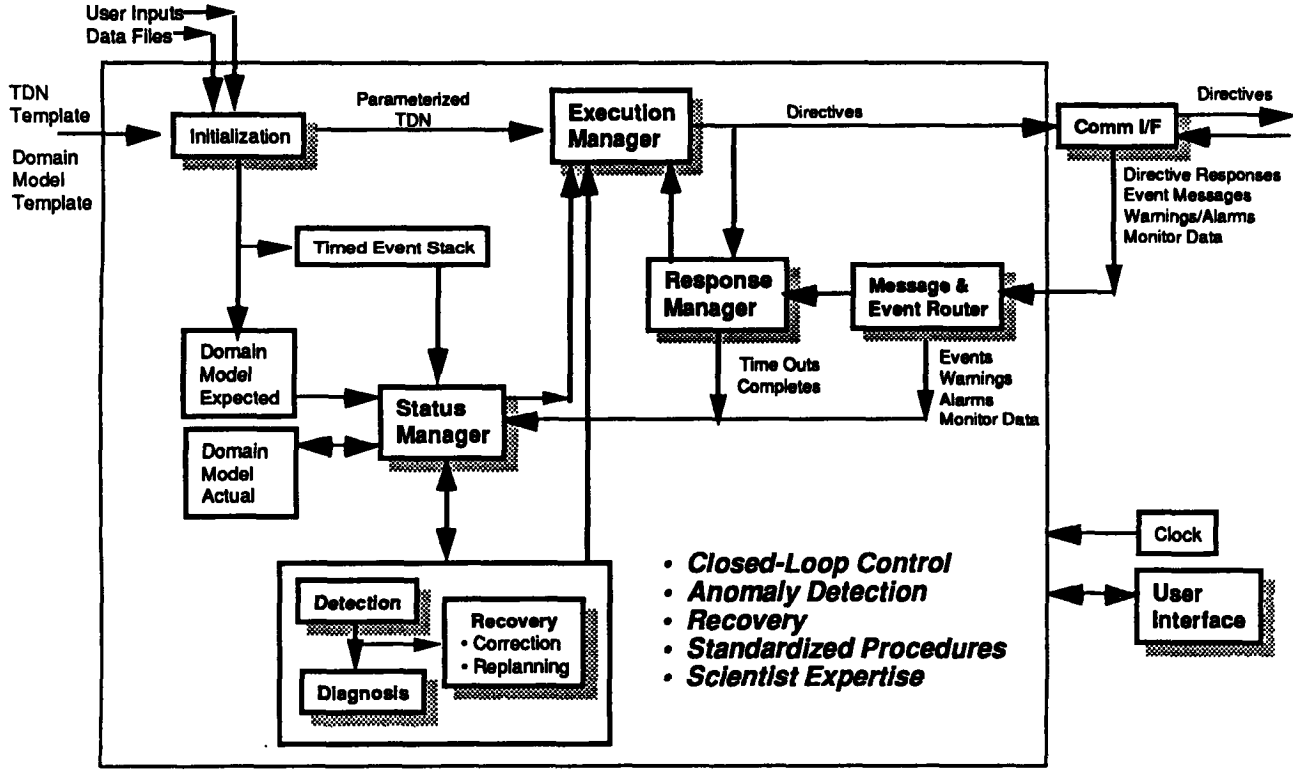


Figure 2. Operator Assistant Architecture

ELECTRONICS

(Session E3/Room B4)

Thursday December 5, 1991

- **Thermoacoustic Refrigeration**
- **Ambient Temperature Recorder**
- **Fiber-Optic Push-Pull Sensor Systems**
- **Commercial Capaciflector**

 PRECEDING PAGE BLANK NOT FILMED

N92-22720

THERMOACOUSTIC REFRIGERATION

Steven L. Garrett and Thomas J. Hofler
Physics Department - Code PH/Gx
Naval Postgraduate School
Monterey, CA 93943

ABSTRACT

A new refrigerator which uses resonant high amplitude sound in inert gases to pump heat will be described and demonstrated. The phasing of the thermoacoustic cycle is provided by thermal conduction. This "natural" phasing allows the entire refrigerator to operate with only one moving part (the loudspeaker diaphragm). The thermoacoustic refrigerator has no sliding seals, requires no lubrication, uses only low-tolerance machined parts, and contains no expensive components. Because the compressor moving mass is typically small (≈ 15 gm) and oscillation frequency is high (≈ 400 Hz), the small amount of vibration is very easily isolated. This low vibration and lack of sliding seals makes thermoacoustic refrigeration an excellent candidate for space applications. Since the thermoacoustic refrigerators use no CFC's and have coefficients-of-performance which are competitive with conventional vapor compression cycle refrigerators, thermoacoustics is also a good candidate for food refrigeration and commercial/residential air conditioning applications. The design, fabrication, and performance of the first practical, autonomous thermoacoustic refrigerator, which will be flown on the Space Shuttle (STS-42), will be described and designs for terrestrial applications will be presented.

INTRODUCTION

The history of refrigeration technology during the second half of the 20th century has been singularly uninteresting. Since the introduction of chlorofluorocarbons (CFC) as the working fluid in a vapor compression refrigeration cycle these chemicals have become dominant in almost all small and medium scale food refrigerator/freezer and building/residential air conditioner applications. That situation is about to change dramatically[1] and, at this moment, unpredictably.

The End of the CFC Era

At present, it is estimated that there is $135 billion of products which uses CFC's within the United States alone[2]. The spectacular success of CFC's in refrigeration was brought about by their good thermodynamic properties (phase changes at modest pressure at the required temperatures) and their excellent chemical stability which made them compatible with hydrocarbon lubricants. It was this compatibility with lubricants which allowed the production of compressors, bathed in oil, which could operate for decades without maintenance. (When was the last time you had to replace your home refrigerator compressor?) It is now known that the chemical stability of the CFC's has lead to their ultimate downfall.

Two recent events are responsible for the "new era" in refrigeration which will dawn at the beginning of the 21st century. The most significant of these is the international ban on the production of CFC's which were found to be destroying the Earth's protective ozone layer. The ban was brought about with the signing on January 1, 1989, of the Montreal Protocols. This international agreement, signed by the United States and thirty-one other industrialized countries, started a stepped reduction in the production of CFC's to 20% of their present levels by 1993, one-half of their present levels by 1998, and imposes a complete ban on their production worldwide by the year 2000.

The second event was the discovery of "high temperature" superconductors and the development of high speed and high density electronic circuits which require active cooling. Although the immediate impact of this emerging requirement for cryocoolers is dwarfed by the revolution which will be brought about by the Montreal

PRECEDING PAGE BLANK NOT FILMED

Protocols, much of the longer term future development of high speed electronics, electro-optics, long-haul fiber-optic communication, and computers, will be dictated by the availability, reliability, and efficiency of cryocooler technology.

Alternative Refrigeration Technologies

The chemical companies have not abandoned the refrigeration business. As one might expect there has been a rush to develop alternative fluids which are not as detrimental to the ozone layer. Thus far, several HCFC compounds, which will be banned in 2010, have been developed but these have exhibited compatibility problems with hydrocarbon lubricants[2] and some have recently been found to be carcinogenic[3]. This has lead others to reconsider fluids which were in use before the rise of CFC (ammonia, carbon dioxide, etc.) and other refrigeration cycles such as Stirling and Malone cycles which do not require phase changes.

It is the purpose of this paper to introduce an entirely new approach to refrigeration which was first discovered[4] in the early 1980's which uses high intensity sound waves to pump heat using inert gases as the working fluid.

THE THERMOACOUSTIC HEAT PUMPING CYCLE

The interaction between acoustics and thermodynamics has been recognized ever since the dispute between Newton and Laplace over whether the speed of sound was determined by the adiabatic or isothermal compressibility of air. Although today there are probably many physicists who might still make the wrong choice (as Newton did!), most physicists have at least been exposed to a lecture demonstration such as the Rijke Tube[5] or have observed (cursed?) Taconis oscillations in liquid helium[6], so they are not surprised that thermal gradients can lead to the production of sound. The reverse process - thermoacoustic heat pumping - is far less well known and was the first intentional demonstration of a new class of intrinsically irreversible heat engines.

Traditional heat engine cycles, such as the Carnot Cycle typically studied in elementary thermodynamics courses, assume that the individual steps in the cycle are reversible. Such analyses, which invoke the First and Second Laws of Thermodynamics, lead to the limiting values for the efficiencies of prime movers and the coefficients-of-performance of refrigerators. These limiting values are never realized in practical heat engines due to the unavoidable irreversibilities, such as thermal diffusion and viscous dissipation, which always reduce performance below the ideal Carnot values. Reversible engines also require various mechanical devices (*eg.*, valves, cams, push-rods, linkages, timing chains, etc.) in order to execute the proper phasing of various cyclic processes (*eg.*, compressions, expansions, regeneration, etc.). In thermoacoustic engines, the irreversibility due to the imperfect (diffusive) thermal contact between the acoustically oscillating working fluid and a stationary second thermodynamic medium provides the required phasing. This "natural phasing" has produced heat engines which require no moving parts other than the self-maintained oscillations of the working fluid.

A Simple, Invisid, Lagrangian Model of the Heat Pumping Process

Although a complete and detailed analysis of the thermoacoustic heat pumping process is well beyond the scope of this study, the following simple, invisid, Lagrangian representation of the cycle contains the essence of the process. A complete analysis[7] would necessarily include the gas viscosity, finite wavelength effects, longitudinal thermal conduction along the stationary second thermodynamic medium and through the gas, and the ratio of the gas and second medium dynamic heat capacities. A schematic diagram of a simple, one-quarter wavelength, $\lambda/4$, thermoacoustic refrigerator is shown in Figure 1. The loudspeaker at the left sets up the standing wave within the gas-filled tube. Its frequency is chosen so that the loudspeaker excites the fundamental ($\lambda/4$) resonance of the tube. The termination at the right-hand end of the tube is rigid so the longitudinal particle velocity at the rigid end is zero (a velocity node) and the acoustical pressure variations are maximum (a pressure antinode). At the loudspeaker end of the tube there is an acoustic pressure node and a particle velocity antinode. To the left of the termination is a stack of plates (the "stack") whose spacing is chosen to be a few thermal penetration depths.

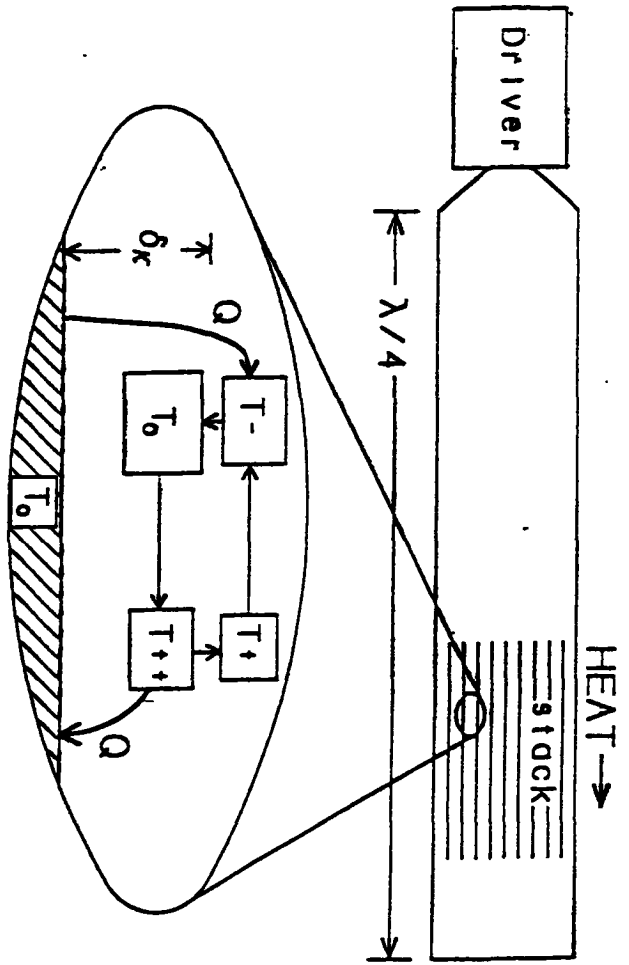


Figure 1. Schematic diagram of a one-quarter wavelength thermoacoustic refrigerator shown in cross-section. The loudspeaker at the left sets up the standing wave in a gas-filled tube. The termination at the right-hand end of the tube is rigid. To the left of the termination is a stack of plates (the "stack") whose spacing is chosen to be a few thermal penetration depths. Below the resonator is an expanded view of a portion of the stack and a parcel of gas undergoing an acoustic oscillation.

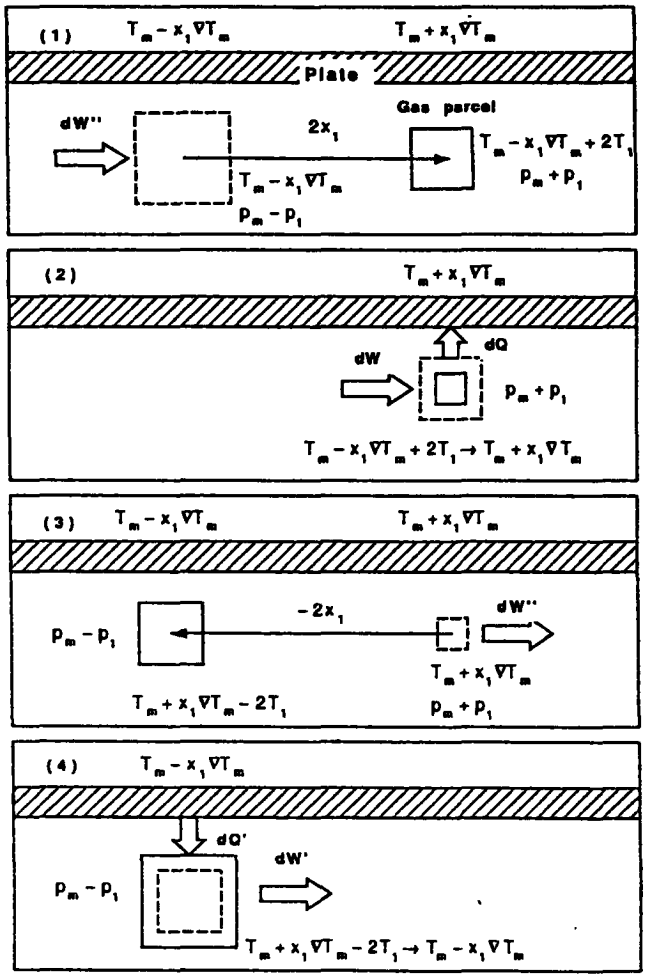


Figure 2. Lagrangian, invisid picture of the thermoacoustic heat pumping cycle. In the first step the gas parcel is adiabatically compressed and its temperature is raised above that of the plate. During the second step the hot parcel rejects some of its heat to the plate and lowers its temperature. The third step is the reverse of the first but since the adiabatic expansion starts from a lower temperature it ends at a lower temperature. During the final step the parcel which is cooler than the plate adsorbs heat from the plate cooling the plate and warming itself.

The thermal penetration depth, δ_κ, represents the distance over which heat will diffuse during a time which is on the order of an acoustic period, $t = 1/f$, where f is the acoustic frequency. It is defined in terms of the thermal conductivity of the gas, κ, the gas density, ρ, and its isobaric specific heat (per unit mass), c_p.

$$\delta_\kappa = \sqrt{\frac{\kappa}{\pi f \rho c_p}} \tag{1}$$

This length scale is crucial to understanding the performance of the thermoacoustic cycle since the diffusive heat transport between the gas and the "stack" is only significant within this region. It is for that reason that the stack and the spacing between its plates are central to the thermoacoustic cycle.

For this analysis we will focus our attention on a small portion of the "stack" and the adjacent gas which is undergoing acoustic oscillations at a distance from the solid stack material which is small enough that a substantial amount of thermal conduction can take place in an amount of time which is on the order of the acoustic period. In the lower half of Figure 1, a small portion of the stack has been magnified and a parcel of gas undergoing an acoustic oscillations is shown. The four steps in the cycle are represented by the four boxes which are shown as moving in a rectangular path for clarity although in reality they, of course, simply oscillate back and forth. As the fluid oscillates back and forth along the plate it undergoes changes in temperature due to the adiabatic compression and expansion resulting from the pressure variations which accompany the standing sound wave. The compressions and expansions of the gas which constitute the sound wave are adiabatic if they are far from the surface of the plate. The relation between the change in gas pressure due to the sound wave, p_1, relative to the mean (ambient) pressure, p_m, and the adiabatic temperature change of the gas due , T_1, to the acoustic pressure change, relative to the mean absolute (Kelvin) temperature, T_m, is given below in equation (2).

$$\frac{T_1}{T_m} = \frac{\gamma - 1}{\gamma} \frac{p_1}{p_m} \tag{2}$$

The polytropic coefficient, γ, is equal to the ratio of the specific heat of the gas at constant pressure to the specific heat at constant volume and is exactly 5/3 for the inert gases. It is smaller for all other gases but is always greater than one.

Although the oscillations in an acoustic heat pump are sinusoidal functions of time, Figure 2 depicts the motion as articulated (a square wave) in order to simplify the explanation. The thermodynamic cycle can be considered as consisting of two reversible adiabatic steps and two irreversible isobaric (constant pressure) steps. The plate is assumed to have a mean temperature , T_m, and a temperature gradient,∇T, referenced to the mean position, $x = 0$. The temperature of the plate at the left-most position of the gas parcels excursion is therefore $T_m - x_1 \nabla T$, and at the right-most excursion is $T_m + x_1 \nabla T$.

In the first step of this four-step cycle, the fluid is transported along the plate by a distance $2x_1$ and is heated by adiabatic compression from a temperature of $T_m - x_1 \nabla T$ to $T_m - x_1 \nabla T + 2T_1$. The adiabatic gas law provides the relationship between the change in gas pressure, p_1, and the associated change in temperature, T_1, as described in equation (2). Because we are considering a heat pump, work, in the form of sound, was done on the gas parcel and it is now a temperature which is higher than that of the plate at its present location.

In the second step, the warmer gas parcel transfers an amount of heat, dQ_{hot}, to the plate by thermal conduction at constant pressure and its temperature decreases to that of the plate, $T_m + x_1 \nabla T$. In the third step, the fluid is transported back along the plate to position $-x_1$ and is cooled by adiabatic expansion to a temperature $T_m + x_1 \nabla T - 2T_1$. This temperature is lower than the original temperature at location $-x_1$, so in the fourth step the gas parcel adsorbs an amount of heat, dQ_{cold}, from the plate thereby raising its temperature back to its original value, $T_m + x_1 \nabla T$.

The net effect of this process is that the system has completed a cycle which has returned it to its original state and an amount of heat, dQ_{cold}, has been transported up a temperature gradient by work done in the form of sound. It should be stressed again at this point that no mechanical devices were used to provide the proper phasing between the mechanical motion and the thermal effects.

If we now consider the full length of the stack as shown in the upper portion of Figure 1, the overall heat pumping process is analogous to a "bucket brigade" in which each set of gas parcels picks up heat from its neighbor to the left at a lower temperature and hands off the heat to its neighbor to the right at a higher temperature. Heat

exchangers are placed at the ends of the stack to absorb the useful heat load at the left-hand (cold) end of the stack and exhaust the heat plus work at the right-hand (hot) end of the stack. The fact that the gas parcels actually move a distance which has typically been on the order of several millimeters means that good physical contact between the heat exchangers and the stack is not crucial since the moving gas provides good thermal contact.

Thermoacoustic Energy Transport

If there were no external (*ie.* useful!) heat load applied to the stack or longitudinal thermal conduction along the stack or through the gas, then eventually the temperature gradient in the plate would approach that caused by the adiabatic processes in the gas. In the absence of gas viscosity, this critical temperature gradient, ∇T_{crit}, is a function only of the gas thermophysical properties, the wavelength of the sound, λ, and the mean position of the stack, x, within the standing wave field.

$$\nabla T_{crit} = \frac{2\pi(\gamma - 1)}{\lambda} \tan\left(\frac{2\pi x}{\lambda}\right) \tag{3}$$

The ratio between the temperature gradient in the stack and the critical gradient, $\Gamma = \nabla T_m / \nabla T_{crit}$, plays an important role in the performance of the stack as will be explained below.

The rate of heat transport or heat pumping power, Q_2, within the stack can be expressed in a simple form if we assume that the stack is much shorter than the wavelength of the sound and we again neglect viscosity.

$$Q_2 = -\frac{\Pi \delta_\kappa}{4} p_1 u_1 (\Gamma - 1) \tag{4}$$

The first term in (4) is simply one quarter of the thermally effective cross-sectional area of the stack, where Π is the stack surface area per unit length and δ_κ is the thermal penetration depth. The heat pumping power is also proportional to the product of the acoustic pressure, p_1, and particle velocity, u_1, in the stack. If the stack were located at a pressure or velocity node, no heat pumping would take place since either the pressure variations which cause the adiabatic temperature changes would be absent or there would be no motion of the gas parcels. Since pressure and particle velocity are proportional (the proportionality constant depends upon the location of the stack within the standing wave field), a doubling of the acoustic pressure would quadruple the heat transport. This is the origin of the subscript "2" on the heat transport symbol, Q_2, which emphasizes the second order dependance of the magnitude of the thermoacoustic heat transport on the square of the acoustic field variables.

The final term in equation (4) is a measure of how close the system is to the limiting temperature gradient. As mentioned before, when $\Gamma = 1$, the gas parcels "see" their adiabatic temperature span on the stack so no heat transfer from the gas to the plate takes place. When $\Gamma = 0$, the temperature of the plate is uniform and a large quantity of heat is pumped by the oscillating gas parcels.

The stack also absorbs work, W_2, at a rate proportional by $(\Gamma - 1)$. The following simple expression for the work absorbed by the stack of length Δx, in the absence of viscosity, can be written if one assumes that the heat capacity of the stack is much greater than that of the gas.

$$W_2 = \frac{\Pi \delta_\kappa}{4} \Delta x \frac{(\gamma - 1)}{\gamma p_m} 2\pi f (p_1)^2 (\Gamma - 1) \tag{5}$$

This work represents the acoustical energy dissipated due to irreversible thermal conduction between the gas and the plate.

The ratio of the heat pumped, Q_2, to the work done, W_2, to pump that heat is defined as the coefficient -of-performance, COP, of the refrigerator. Since the temperature spanned by the stack, $\Delta T = \Gamma \nabla T_{crit} \Delta x$, one can show that the thermoacoustic $COP = \Gamma \, COP_{Carnot}$, where $COP_{Carnot} = \Delta T / T_{hot}$, is the ideal coefficient-of-performance dictated by the First and Second Laws of Thermodynamics. Since for a heat pump, $\Gamma < 1$, we see that the thermoacoustic COP is always less than that of Carnot. We also see that with the thermoacoustic heat pump, there will be the same competition between power density and efficiency which exists in all heat engines. As we pointed out earlier, there is no useful heat pumped when $\Gamma = 1$, which is the point where the efficiency is at its maximum.

The general results derived above from the simple, invisid picture are essentially preserved when the viscosity of the gas is included but the detailed mathematical descriptions are substantially more complicated. Discussion of these equations is well beyond the scope of this introduction to thermoacoustics. The reader is referred to the excellent review article by Dr. Swift [7] for a detailed derivation and discussion of this more complete analysis.

THE SPACE THERMOACOUSTIC REFRIGERATOR (STAR)

STAR was the first attempt to exploit the advantages of the thermoacoustic heat pumping cycle for cryocooler applications in space. It is intended to operate autonomously in low Earth orbit aboard the Space Shuttle in a Get Away Special (GAS) canister. It derives its power from an internal battery power source (700 Watt-hours) and was optimized for a modest temperature span ($\Delta T \leq 80$ °K) and small heat loads ($Q_{cold} \leq 5$ Watts). Due to requirements of small size and light weight imposed by the GAS envelope, it operates as a frequency of about 400 Hz and is driven by an electrodynamic loudspeaker[8]. Its frequency of operation is adjusted automatically[9] to keep the system on resonance. The resonator length is approximately equal to a quarter wavelength of sound in a mixture of helium and argon or helium and xenon gas[10] maintained at a mean pressure of ten atmospheres (150 psia). It has a single stack with a uniform spacing made of polyester film (Mylar™) and fishing line[10,11] which is spiral wound like a "jelly roll". The stack used in STAR is 7.9 cm in length and 3.8 cm in diameter. Copper-strip, parallel plate heat exchangers are located at either end of the stack. Figures 3 are scale drawings which show the details of the acoustical sub-systems.

Acoustical Sub-Systems.

The driver housing does more than simply hold the electrodynamic driver. The considerable mass and size of the housing is a consequence of the fact that a commercially available loudspeaker was modified for this application. The driver housing serves as a heat sink for the heat generated by the resistive losses in the driver voice coil and the heat pumped away from the cold end of the resonator. It also contains the ten atmospheres of the helium/xenon gas mixture which is the "working fluid" within the refrigerator. The housing is bolted to the standard 12 inch bolt circle provided on the GAS cannister lid which acts as the radiator while on orbit.

The driver voice coil is attached to an aluminum reducer cone which is in turn attached to a nickel electroformed bellows. The bellows provides a means of transferring acoustic pressure to the resonator without the need for sliding seals. A miniature accelerometer is attached to the surface of the reducer cone opposite the bellows to monitor the displacement magnitude and its phase relative to the acoustic pressure at the face of the bellows. That acoustic pressure is monitored by a piezoelectric quartz microphone[12] followed by a MOSFET impedance converter located within the driver housing in close proximity to the microphone.

The resonator is a modified quarter wavelength tube. The open end is terminated by a tapered "trumpet" and sealed by a surrounding sphere. Thus, an "open" termination is simulated while still allowing the resonator to retain the ten atmosphere gas mixture. The thermoacoustic stack and heat exchangers are located in a section of the resonator designed to allow a minimum of heat conduction back to the cold end. The resonator is wrapped with multiple layers of superinsulation to prevent heating by thermal radiation. An electrical strip heated element is attached to the cold end of the resonator near the cold heat exchanger to permit measurement of refrigerator performance with a variable and quantifiable heat load. A thermal isolation vacuum chamber surrounds the resonator and seals against the bottom surface of the driver housing with an O-ring.

Electrical Sub-Systems.

In order for the refrigerator to operate autonomously in space, a family of analog and digital electronic circuits are employed to monitor the "health" of the system, keep the driver running at the proper amplitude and frequency, and to acquire and store useful data for post-flight analysis. The design and function of each of these sub-systems is documented in Reference [9].

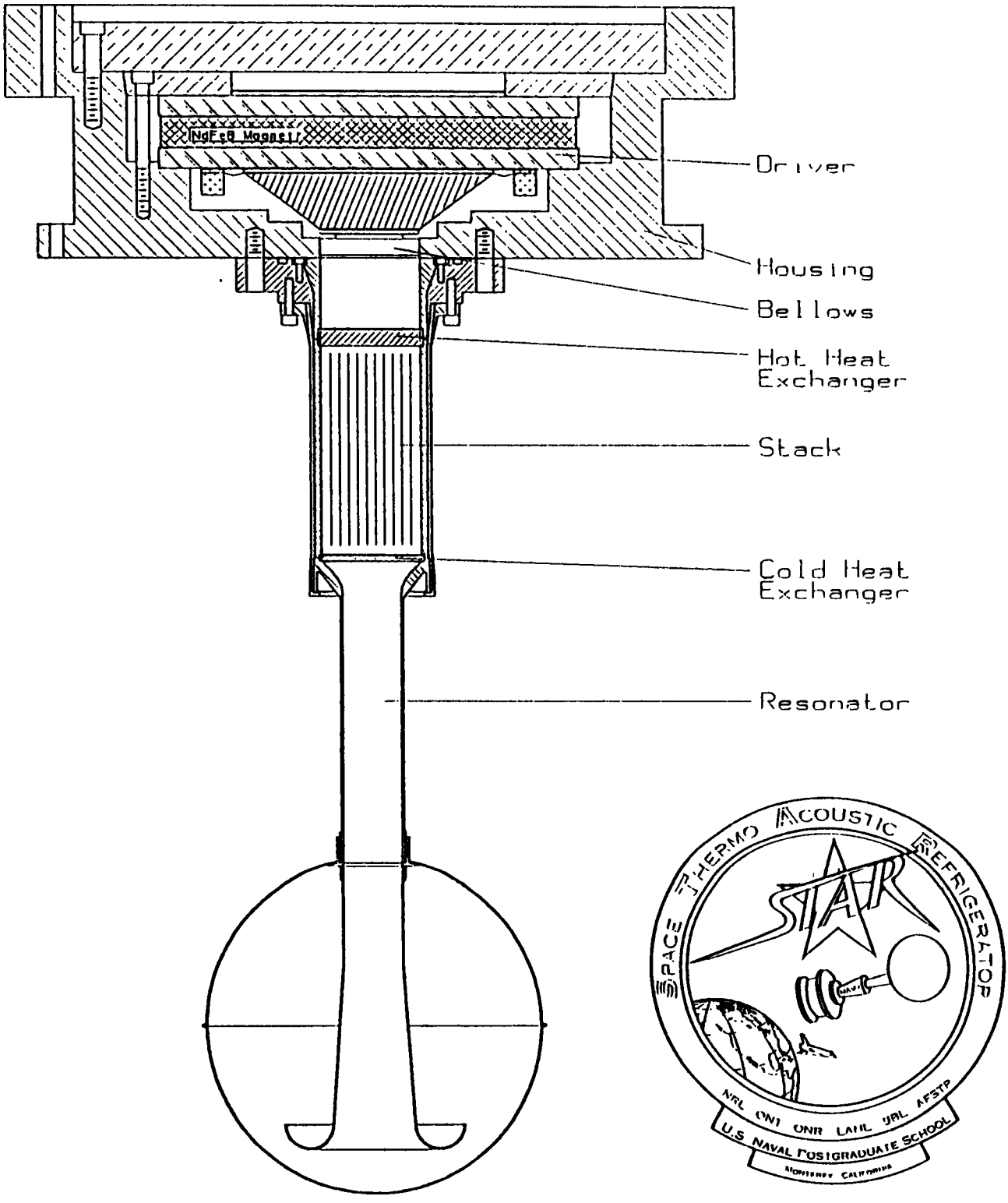


Figure 3. Scale drawing of the STAR acoustical sub-systems showning the driver, resonator, stack, and heat exchangers. Shown to the right of the sphere is the official logo for the experiment which will be flown on Space Shuttle mission STS-42 scheduled for launch on 22 January 1992.

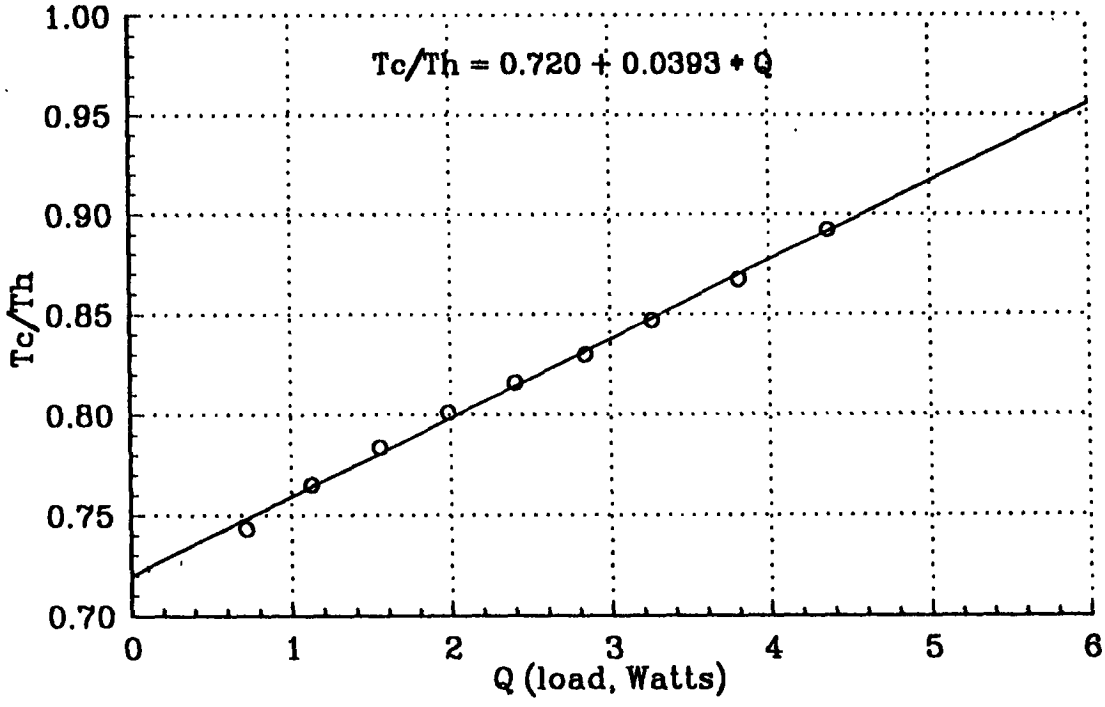


Figure 4. The ratio of the cold and hot heat exchanger temperatures, T_c/T_h, as a function of heat load on the cold heat exchanger at an acoustic pressure ratio of $p_1/p_m = 2\%$ in a lean mixture (4% Xe) of helium and xenon.

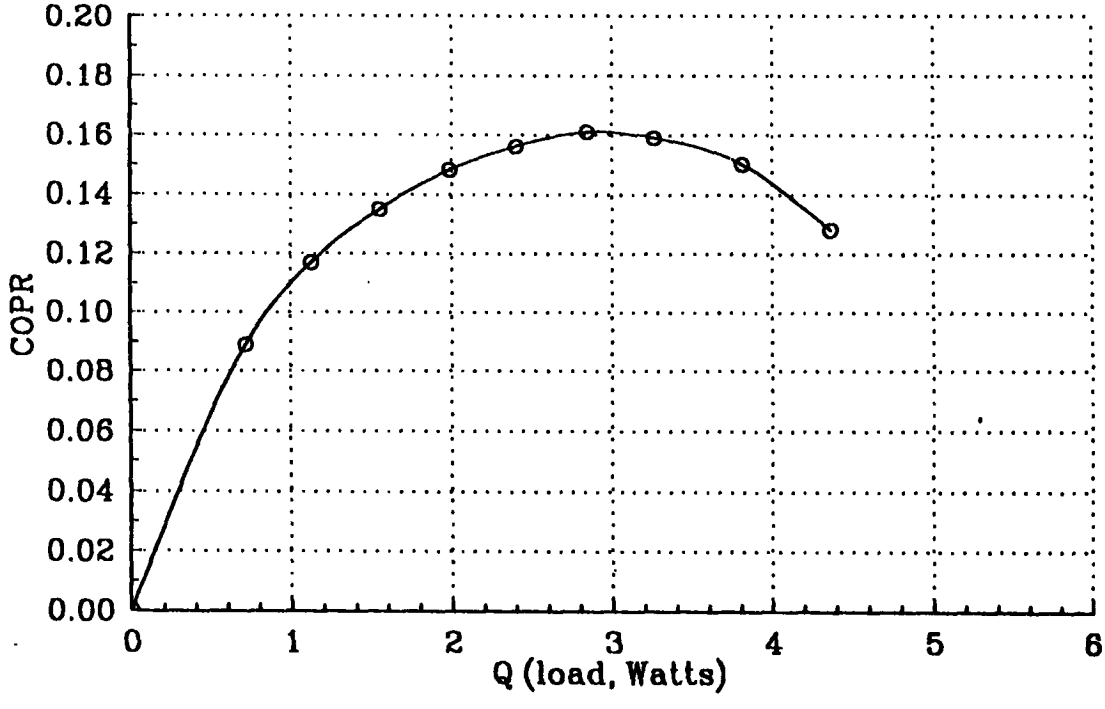


Figure 5. Measured coefficient-of-performance relative to the ideal Carnot coefficient-of-performance (COPR) as a function of heat load for STAR operating at an acoustic pressure ratio of $p_1/p_m = 2\%$ in a lean mixture (4% Xe) of helium and xenon.

The circuit which is unique to this application is the Resonance Control Board (RCB). Its function is to maintain the system at acoustical resonance and to control the amplitude of the acoustical pressure at pre-determined levels dictated by the Controller Board. As the temperature changes, the sound speed changes, hence, for fixed resonator dimensions, the resonance frequency will be a function of temperature.

The RCB maintains of the resonance condition by comparing the relative phase of the microphone and accelerometer outputs. At resonance the pressure and velocity of the gas mixture at the bellows location must be in-phase. Therefore, the pressure and acceleration are required to be in quadrature (*ie.* 90° out-of-phase). The two signals are electronically multiplied together and the dc-component of their product, which is proportional to the cosine of their phase difference, is used as an error voltage. This error voltage is electronically integrated and fed back to a voltage controlled oscillator to close the phase accurate phase-locked-loop circuit which maintains the entire system at resonance.

Refrigerator Performance.

The performance of STAR can be summarized by the two graphs which are presented here as Figures 4 and 5. Other measurements of the performance of the entire system, including the electrodynamic driver are included in Reference [11]. The STAR performance shown in Figures 4 and 5 is by no means the best which has been achieved in a thermoacoustic refrigerator of this style. Other improvements to the design, such as the use of a stack which has non-uniform spacing, has produced a single-stage, no-load temperature span of 118 °K even without the use of gas mixtures to reduce viscous losses or improve the coefficient of performance. COPR's as large as 20% have also been measured in a similar refrigerator over the same temperature span[10].

HIGH POWER THERMOACOUSTIC REFRIGERATORS

The STAR is the first in a series of space cryocoolers now under development at the Naval Postgraduate School. Those applications generally require relatively little heat pumping power (a few watts) but a very large temperature span (ΔT = 100 to 200 °K). The requirements of residential refrigerator/freezers and air conditioners are opposite of those for spacecraft cryocoolers. Those application require modest temperature spans (ΔT ≈ 25 - 45 °K), but much higher heat pumping powers on the order of hundreds of watts for refrigerators and thousands of watts for air conditioners. They also are typically powered by 110 volt alternating current rather than 28 volt direct current. There are several design modifications which are therefore required to adapt this established space cryocooler technology to residential applications, but these modifications do not present any substantial technological barriers.

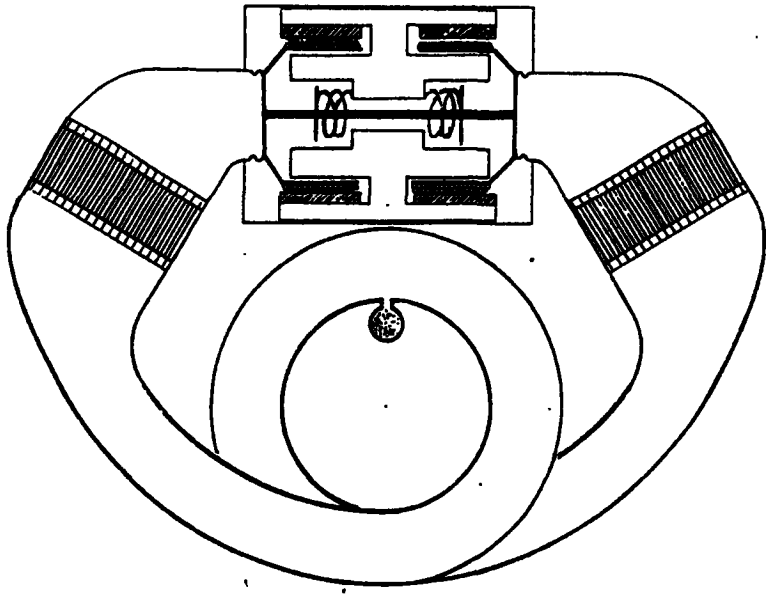


Figure 6. Schematic diagram of a half-ton (2 kW) capacity thermoacoustic chiller. The half-wavelength resonator operates at 60 Hz, is powered by a single double-acting electrodynamic driver and contains two "stacks".

Operation at fixed frequency (60 Hz or 120 Hz) can be accomplished by selective absorption of one component of the gas mixture as the temperature of the working fluid changes. Half-wavelength resonators, such as the one shown in Figure 6, can utilize both oscillating faces of a single driver which can be more massive and hence more efficient because it operates at fixed frequency. The presence of two stacks reduces the required heat pumping capacity of each individual stack although the diameters of the stacks would necessarily be larger. The design shown in Figure 7 has stack diameters of 18 cm, operates at 60 Hz, is about 90 cm wide, 60 cm tall, and 20 cm thick. It should be capable of pumping about 2 kW of heat across a 30 °C temperature span with a COP ≈ 3-4 (about 30% of Carnot performance) including electroacoustic conversion efficiencies. At the present time, the greatest uncertainty in the use of thermoacoustic engines in high power cooling applications is the design of the heat exchangers.

ACKNOWLEDGEMENTS

The authors are indebted to Dr. Greg W. Swift, Los Alamos National Labs, for several conversations regarding this new technology and his excellent review article on thermoacoustic engines whose contents was exploited extensively in the first part of this paper[7]. The development of the Space Thermoacoustic Refrigerator was the work of several staff and students at the Naval Postgraduate School. Staff members include David Rigmaiden (Payload Manager), Jay Adeff (Physicist), Glenn Harrell (Machinist), Jim Horning and Ron Phelps (Programmers) and CDR (Dr.) David Gardner, NOAA Corp. (Group Leader). The students who worked on the project as part of their Master's degrees were LT Michelle Fitzpatrick, USCG (Driver), LT. Michael Susalla, USN (Thermodynamics and Gas Mixtures), CAPT David Harris, Canadian Forces (Driver), LT Richard Volkert, USN (Driver), and MAJ Ronald Byrnes, USA (Electronics). The loudspeakers used as the drivers were donated by Harmon-JBL and modified by their engineer, Fancher Murray. This work is supported by the the Office of Naval Research (Physics Division), the Office of Naval Technology and the Naval Postgraduate School Direct Funded Research Program. Substantial funding and technical support for STAR was provided by the Naval Research Laboratory - Spacecraft Engineering Division. The payload launch services were provided by the Air Force Space Test Program.

REFERENCES

1. S. Derra, "CFC's: No Easy Solutions," Res. & Dev. 32(5), 56-66 (1990)
2. R. Monastersky, "Decline of the CFC Empire," Sci. News 133, 234-236 (1988).
3. A. K. Naj, "CFC Substitute," *Wall St. Journal*, (2 July 1991).
4. J. C. Wheatley, T. Hofler, G. W. Swift, and A. Migliori, "Experiments with an Intrinsically Irreversible Acoustic Heat Engine," Phys. Rev. Lett. 50, 499 (1983); "Intrinsically irreversible heat engine," U. S. Patent No. 4,489,553 (Dec. 25, 1984).
5. J. W. Strutt (Lord Rayleigh), *The Theory of Sound*, 2nd ed., Vol. II (Dover, 1945), §322j.
6. T. Yazaki, A. Tominaga, and Y. Narahara, "Experiments on thermally driven acoustic oscillations of gaseous helium," J. Low Temp. Phys, 41, 45 (1980).
7. G. W. Swift, "Thermoacoustic Engines," J. Acoust. Soc. Am. 84(4), 1145-1180 (1988).
8. D. A. Harris and R. E. Volkert, "Design and Calibration of an Electrodynamic Driver for the Space Thermoacoustic Refrigerator", Master's Thesis, Naval Postgraduate School, Monterey, CA (Dec. 1989); DTIC Report No. AD A212 022.
9. R. B. Byrnes, Jr., "Electronics for Autonomous Measurement and Control of a Thermoacoustic Refrigerator in a Space Environment", Master's Thesis, Naval Postgraduate School, Monterey, CA (Dec. 1989); DTIC Report No. AD B141 388.
10. M. P. Susalla, "Thermodynamic Improvements for the Space Thermoacoustic Refrigerator", Master's Thesis, Naval Postgraduate School, Monterey, CA (June, 1988); DTIC Report No. AD A196 958.
11. J. A. Adeff, "Measurement of the Space Thermoacoustic Refrigerator Performance", Master's Thesis, Naval Postgraduate School, Monterey, CA (Sept. 1990).
12. T. Hofler, "Accurate Acoustic Power Measurements with a High-Intensity Driver," J. Acoust. Soc. Am. 83, 777 (1988).

N92-22721

AMBIENT TEMPERATURE RECORDER*

Larry D. Russell
Electronic Systems Branch, N213-2
NASA Ames Research Center
Moffett Field, California

ABSTRACT

A temperature data recorder, designated the Ambient Temperature Recorder or ATR-4, has been developed at NASA Ames Research Center to meet particular requirements for space life sciences experiments. The small, self-contained, four-channel, battery-powered device records 32 kilobytes of temperature data over a range of -40 to +60 deg. C at four sampling intervals ranging from 1.875 to 15 min. Data is stored in its internal electronic memory for later readout by a personal computer.

The ATR-4 has been used to record external instrument temperatures on the US-USSR Cosmos Biosatellite, animal enclosures and other space life sciences experiments on several recent Space Shuttle flights. It also has been used to record instrument temperatures on high altitude aircraft flight experiments.

Commercial potentials include the variety of needs for a small, self-contained, unattended temperature recorder for the transportation of perishables or for recording life system or process temperatures over a period of time from minutes to months.

INTRODUCTION

The Ambient Temperature Recorder, ATR-4, measures temperature on one to four channels at pre-selected sampling rates and stores the data for subsequent readout by a personal computer (PC). The ATR-4 was designed to meet the recurring need of NASA Space Life Sciences' experiments for a small, reliable, flight-qualified, and easy-to-use temperature recorder. Operational requirements included a temperature recording range of -40 to +60 deg. C, sampling intervals from 1.875 to 15 min., sufficient memory for typical 10-day missions, and small size (less than 100 x 60 x 25 mm). The unit can be modified to measure temperature from -60 to +155 deg. C, and record data at intervals from 7.0 sec. to 15 min.

The ATR-4 system consists of the ATR-4 recorder, a Computer Interface Unit and operating software. The ATR-4 has been qualified for Space Shuttle use, which includes acceptance testing, verification, and Flight and Ground Phase III Safety Data packages per National Space Transportation System requirements (1).

DESCRIPTION

The ATR-4 measures and records temperatures from -40 to +60 deg. C with an accuracy of +/-1 deg. C. The temperatures are converted to 8-bit digital values, which provides a resolution of 0.4 deg. C over the 100 deg. C temperature range. Up to four external temperature sensor probes may be attached. The recorder's operating range is the same as its measurement range, and the ATR-4 may be used to record its ambient temperature (without probes) by selecting the internal temperature sensor which is thermally bonded to the inside of the front cover of its case.

---

* Presented at the 37th International Instrumentation Symposium, Instrument Society of America, May 9, 1991.

The ATR-4 weighs 135 grams and is contained in an anodized aluminum case measuring 23 x 41 x 86 mm which was designed to conform to the size and shape of the small Spacelab Biorack specimen container (Figure 1). Over-center latches secure the o-ring sealed front cover which contains the external probe receptacles. The front cover also supports the entire electronics assembly and batteries on a single circuit board (Figure 2). Two 3.5 volt lithium batteries supply power for the circuit and RAM memory. The battery life is conservatively rated for one year, but its actual shelf or operating life is approximately two years.

A crystal controlled oscillator establishes timing for the circuitry and the time interval between measurements. Four time intervals (1.875, 3.75, 7.5 and 15 min.) may be selected by the Rate switch inside the ATR-4 case. A Probe switch selects the number of probes used, and the Internal/External switch selects between the internal sensor or an external probe for channel 1. There are two other switches - the Reset and the Stop/Operate Jumper switch - described below. All switches are manually set. Figure 3. illustrates the ATR-4 layout.

Digital temperature data is stored sequentially in a 32K byte RAM, and is transferred serially to a PC for readout via a Computer Interface Unit built for the purpose. Connection to the ATR-4 is made with a miniature 20-pin rectangular connector and ribbon cable from the Computer Interface. The Computer Interface Unit is connected by standard RS-232 serial connection to an IBM-compatible computer using the MS-DOS operating system. Figure 4. shows the ATR-4 connected to the Computer Interface Unit and a readout computer.

The ATR-4 has two setup methods. Normal operation of the ATR-4 uses a computer to assist setup and provide storage of setup parameters and experiment information in a Header at the beginning of the ATR-4 memory. With the ATR-4 connected to a computer through the Computer Interface Unit, the ATR Operating Program, guides the user through setting of the ATR-4 probe and sampling interval switches and monitors the switch positions to confirm the desired settings. Either an "immediate start" or "delayed start" of the ATR-4 may be chosen by setting the Stop/Operate Jumper switch to the proper position. In the stop position the timing oscillator is shut down, and no data is taken. In the operate position, timing starts and temperature sampling begins.

The second method is manual setup of the ATR-4 without a computer. It consists of setting the switches and pressing the Reset button, which resets the RAM address to zero and causes data to be recorded, starting at the beginning of memory. Manual setup does not allow a Header to be stored at the beginning of memory.

Temperature data is stored sequentially in memory. There is no off-on power switch, and the memory is never erased except by command as an option. Normally, each time the recorder is set up, the memory is reset and previous data is overwritten. When the memory is filled with data, the ATR-4 stops recording and enters a standby state. ATR-4 recording duration ranges from a minimum of 10 days for 4 probes at the 1.875 min. sampling interval to a maximum of 342 days for 1 channel at the 15 min. interval.

Data readout is accomplished by connecting the ATR-4 to the Computer Interface Unit and following the on-screen directions of the Readout section of the ATR Operating Program. At the beginning of readout, the computer stores an end-of-data (EOD) mark in memory immediately following the last recorded data. It then resets the ATR-4 and reads out the Header and data up to the EOD mark. The transferred data is stored in a file in the computer for subsequent tabular and graphic presentation.

ATR-4 CIRCUIT AND COMPONENTS

Referring to the simplified circuit schematic, Figure 5., the ATR-4 consists of an oscillator-counter timing section, a signal input analog section with operational amplifiers and an analog-to-digital converter (ADC), and a RAM memory section.

The integrated circuits (IC) are low power 4000 series CMOS. To conserve battery energy, the analog input and ADC circuitry, which draw 9 ma, are normally unpowered and only switched on at each sampling interval with a low forward-drop P-MOSFET transistor switch. The remaining digital IC's and RAM draw approximately 30 microamps when in standby mode between sampling intervals.

Temperature is measured with an Analog Devices AD590 integrated circuit temperature sensor that provides a current output linearly proportional to the absolute temperature. A negative 5 volts is applied to the AD590 by a 7660 inverter which produces a negative temperature current that is summed with a positive offset current from a

REF43 precision 2.5 volt reference, which sets the low temperature measurement point. A quad TLC27L9 operational amplifier with feedback resistor-controlled gain inverts the negative input, and produces a 0-2.5v signal output. Precision (0.05%), low temperature coefficient (5 ppm/deg. C) resistors are used to set the offset current, which establishes the -40 deg. C low point, and the gain, which establishes the 100 deg. C measurement range. With the precision resistors, calibration is not required for the specified +/-1 deg. C accuracy. The 0-2.5v analog temperature signals are then converted to 8-bit bytes by a MAX154 four channel ADC.

By selecting different offset and feedback resistor values the low temperature point may be set as low as -60 deg. C, and the range can be adjusted for a high temp point up to 150 deg. C. The narrower the range the higher the resolution. A 100 deg. C range results in a 0.4 deg. C resolution. The operating temperature range of the ATR-4 circuit is limited to -60 to +85 deg. C.

Circuit timing is based on a crystal controlled 4536 programmable timer that provides a periodic sampling pulse at 1.875, 3.75, 7.5 and 15 min. intervals, depending on the Rate switch selection. There are four additional intervals of 7, 14, 28 and 56 seconds which are excluded from ATR-4 use by a mechanical stop on the switch. The sampling pulse sets a NOR gate latch which powers the analog and ADC circuits through the P-MOSFET transistor switch and starts a 4040 circuit timing counter which sequences the ADC read and write and the RAM storage operations through various combinatorial logic gates. Other 4040 output stages are combined through the Probe switch to send a reset pulse to the sampling latch which determines how many channels will be converted. Conversion requires about 3 microseconds, and the time between channel conversions is 6.9 milliseconds.

Data bytes are stored in a 32Kx8 static RAM which is sequentially addressed by another 4040 counter with spill-over into one half of a 4520 counter. The other half of the 4520 provides channel sequencing for the ADC. When the RAM is filled, the next counter stage disables the 4536 oscillator and the ATR-4 reverts to the 30 microamp low power standby state. The oscillator may also be disabled by setting the Stop/Operate Jumper switch to the Stop position.

The ATR-4 switch monitoring feature is implemented by connecting the logic level of each switch position in an 8-bit status byte that is stored in a HC244 register and monitored by the setup- readout computer. The ATR-4's serial number also is read by the computer to identify the unit. A serial number from 0 to 255 is assigned to each ATR-4 by connecting each of its eight data lines to either Vc or ground with 1M resistors. This binary coded data bus resistance pattern is interpreted by the computer as the serial number.

Power for the circuit is provided by two series-connected, hermetically-sealed lithium thionyl chloride 3.5v batteries through a 2936 low forward-drop 5v regulator. Dual redundant reverse-charge blocking diodes and individual battery fuses are provided to meet Space Shuttle battery safety requirements (2). There is no off/on power switch. The low power static RAM and CMOS digital IC's are always on at a low 30 microamp level to insure memory retention.

Besides the eight data lines plus ground, there are six control lines and four analog signal lines connected from the circuit to the 20-pin connector. The control lines are: oscillator disable, analog on, clock, read/write, switch enable and reset. They are used by the Computer Interface Unit for setup and readout functions. The four 0-2.5v analog temperature signals are available at the connector for testing purposes, and can be read by an ancillary hand-held Field Tester which was built for optional use at remote sites where no computer was available. The Field Tester also indicates the state of the stop/operate and manual reset switches. Control line inputs and data lines are protected by 10K current-limiting resistors, and signal inputs are protected by 1K resistors along with rail-to-rail shunting diodes.

ATR-4 ASSEMBLY AND PRODUCTION

To meet small package constraints, the ATR-4 was assembled using surface mount components. Military standard (MS) grade components were used wherever possible for reliability. Surface mount ICs, however, were not available in MS grade in the preferred small outline (SOIC) package with its gull-wing shaped leads. The common MS package was either a leadless chip carrier, which did not allow for stress relieving compliance when soldered to a circuit board, or a J-leaded package with leads bent beneath the chip which prevents adequate inspection. Industrial rated, plastic encapsulated gull-wing SOIC's with typical operating range of -40 to +85 deg.

C were used. For the -40 to +60 deg. C thermal cycle operating condition of the ATR-4, deterioration of the plastic encapsulated SOIC's will be a factor limiting its lifetime to an anticipated 15 years with occasional usage.

A single multilayer printed circuit (PCB) board having six circuit layers was required. Polyimide PCB material was used for its low thermal expansion. The small resistor, capacitor and diode chip components were mounted on the bottom of the board, and the IC's and through-hole switch and connector components were mounted on top.

The initial ATR-4's were assembled using manual placement and a surface mount hot air rework station to solder the components. A small production run was done with a simple table top vapor phase soldering unit (Vaporette) which allowed all of the surface mount components to be soldered in two passes, and subsequent production has been done with a larger vapor phase soldering machine. The small chip components were positioned and soldered on the bottom of the board first (bottom side up), and then the IC's were placed and soldered on the top side. Solder paste was dispensed using a squeegee and a stainless steel solder mask. Component adhesive was not required. The boards were cooled promptly after solder reflow to enhance small grain structure and stronger joints in the solder alloy (3). Final assembly of the units included applying silicone conformal coating for protection from contamination.

COMPUTER INTERFACE UNIT

The ATR-4 is set up and read out by an IBM XT/AT-compatible computer through an AC line-powered Computer Interface Unit. The Computer Interface Unit translates the 8-bit parallel data and decodes the control lines into serial data for downloading via the computer's RS-232 serial port. Data transfer is at 9600 baud.

The Computer Interface Unit is based on the Cybernetics CY233 network controller chip which provides parallel-serial data conversion and also decodes the ATR-4 control lines. A single circuit board contains the controller chip, associated logic chips and power supply.

SOFTWARE

The ATR-4 software is written in PASCAL and provided on a diskette as a single operating program labeled ATR. The program is entirely menu driven and consists of the following functions: Setup, Readout, Data Display, and Utilities.

SETUP is used to prepare the recorder for recording data. It guides the user in setting the Rate and Probe switches and entering Header information. The Header information includes the ATR-4 serial no. (read automatically), setup and battery dates, probe numbers, and a comment field. The ATR-4 switch settings are monitored by the program to minimize setup error and are automatically recorded in the Header.

READOUT enables temperature data stored in the ATR's memory to be transferred to a PC. The end of current data is identified by an end of data (EOD) marker written by the computer immediately after the last current data byte in the ATR-4 memory. The ATR-4 is then reset by the computer and the memory is read out sequentially up to the EOD marker. The temperature data is not affected by readout and may be re-read out indefinitely. Old data is not erased. It is retained in memory until subsequent reset and recording writes over the previous data.

DATA DISPLAY includes tabular and graphic presentations. Tabular data is listed, along with the elapsed time, for the number of probes used. A printed copy of the tabular data, including header information, may be selected.

The graphic display plots temperature versus elapsed time from the start of recording. The program allows the user to select a particular time segment of data, temperature range for plotting, and the number of probes to be displayed. The elapsed time may be correlated to another time, such as mission or actual time, by specifying the start time. Graphic data can be printed with an Epson compatible dot matrix printer.

UTILITIES include selection of the computer serial port (COM1 or COM2), the disk drive to be used for data, an optional ASCII text data file for use with other data reduction programs, and an ATR-4 memory test. The Memory Test writes and reads a test byte to all locations of memory. Its use, of course, writes over all data existing in the ATR-4. Data may be erased with a Memory Erase selection which writes a zero value to all memory locations.

RQA AND ACCEPTANCE TESTING

Reliability and quality assurance (RQA) procedures included failure modes analysis, parts screening, assembly certification and acceptance testing. Failure modes analysis resulted in elimination of an off-on switch. Assembly certification included a 55 cycle temperature test from -50C to +70C (10 degrees beyond each end of the ATR-4's operating range) of the Vaporette-formed surface mount solder joints.

To meet Space Shuttle flight requirements (4,5), the ATR-4 was subjected to offgas and electromagnetic emission testing, and all units were subjected to acceptance testing. The acceptance tests included a 100 hour burnin at +62C, pressure differential test (less than 200 Torr vacuum environment), fluid resistance test (water submersion), random vibration tests, and thermal cycling. The thermal cycle tests included five cycles with the ATR-4 operating over a -38 to +58 deg. C range. The random vibration test applied a sweep and composite of sinusoidal vibrations to each axis of the ATR-4. After each segment of the acceptance testing, limited or full functional operation tests were performed. The full functional test included a full range accuracy test to confirm the +/-1C specified accuracy.

APPLICATIONS

An early version of the ATR-4 was flown on the Russian Cosmos 2044 Biosatellite in 1989 to record external surface radiation dosimeter temperatures as part of a joint US-USSR program. The ATR-4 has flown on several recent Shuttle flights for recording temperatures in various experiments: a Growth Hormone Concentration and Distribution in Plants experiment on STS-34, a Circadian Rhythm of Neuraspora experiment on STS-32, and a Physiological Systems Experiment with Animal Enclosure Modules on STS-41.

A derivative of the ATR-4, called the Flight Temperature Recorder (FTR), is currently used on NASA ER2 high altitude earth resources research aircraft to record instrument temperatures, including values below -60 deg. C on ozone detection instrumentation.

SUMMARY

A small four-channel, battery powered ambient temperature recorder (ATR-4), has been developed to meet NASA Life Science Shuttle payload requirements. It measures temperatures from -40C to +60C at selected intervals, and stores up to 32K bytes of data for subsequent readout by a personal computer.

ACKNOWLEDGEMENTS

Many individuals have contributed to the development of the ATR-4 recorder. In particular, program leadership was provided by J. P. Connolly, design contributions were made by W. F. Barrows, software by G. S. Buhtz and D. J. Ungar, and RQA by S. Askarinam. ATR-4 assembly was performed by the Ames Electronic Instrument Services Branch.

REFERENCES

1. "Safety Policy and Requirements for Payloads Using the Space Transportation System (STS)," NASA, NTST 1700.7B.

2. "Manned Space Vehicle Battery Safety Handbook," NASA, JSC-20793.

3. Devore, John A., "The Makeup of a Surface Mount Solder Joint", Circuits Manufacturing, Vol. 30, No. 6, June 1990.

4. "Shuttle/Payload Interface Definition Document for Middeck Accommodation," NASA, NSTS 2100-IDD-MDK.

5. "Spacelab Payloads Accommodation Handbook (SPAH)," NASA, SLP/2104.

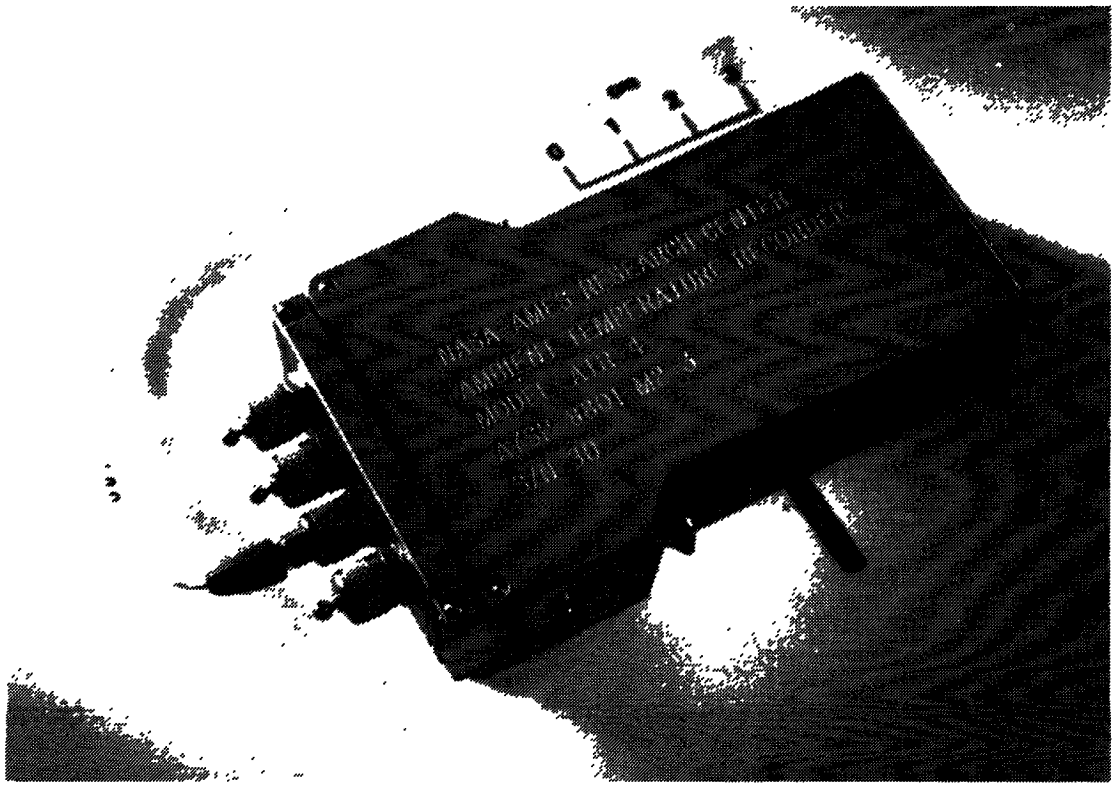


Figure 1. Ambient Temperature Recorder ATR-4

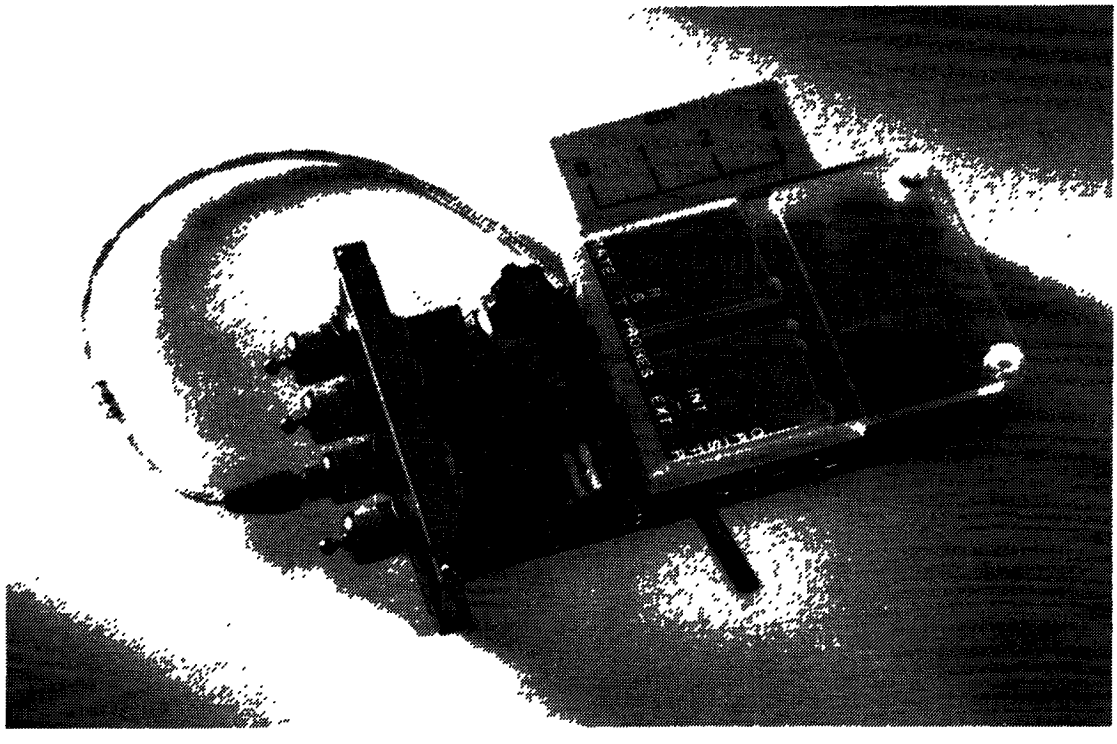


Figure 2. ATR-4 Circuit Board

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

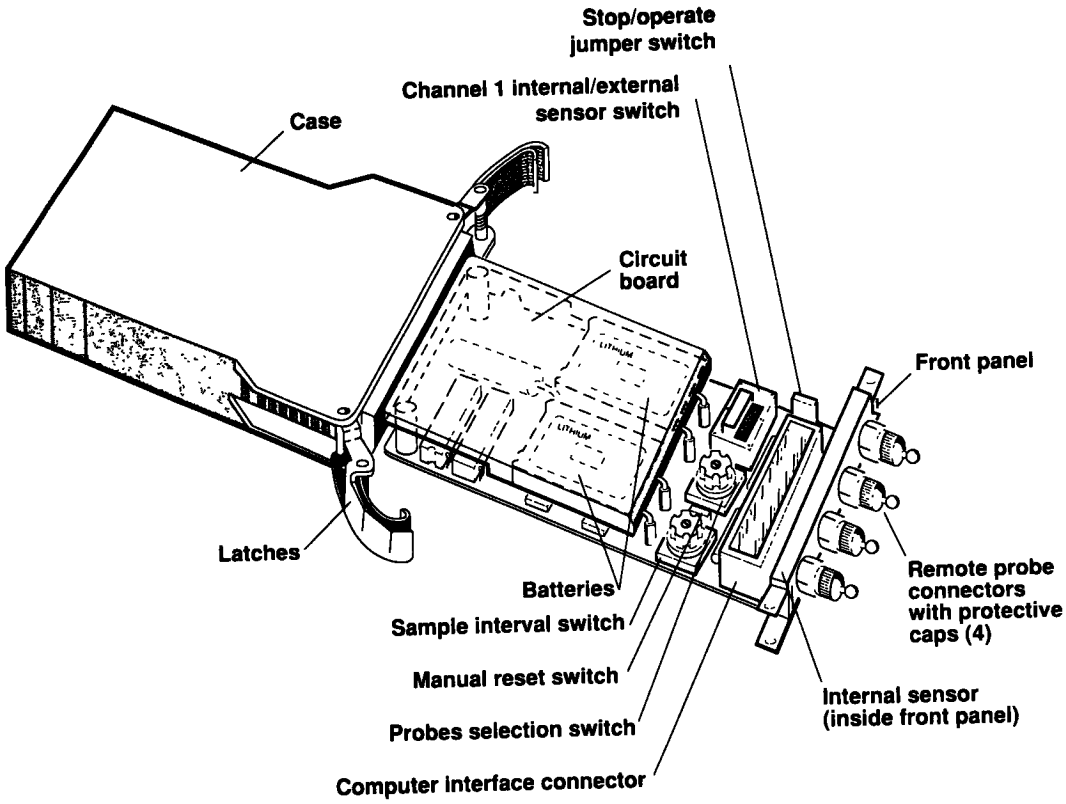


Figure 3. ATR-4 Layout

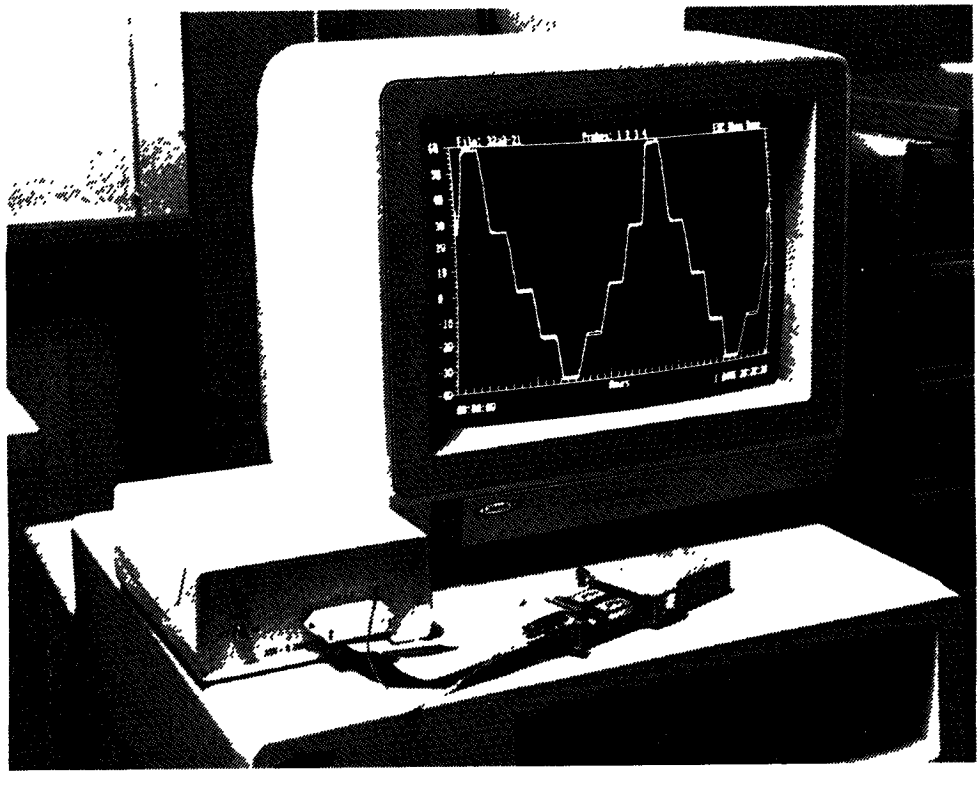


Figure 4. ATR-4 Connected to Computer Interface Unit

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

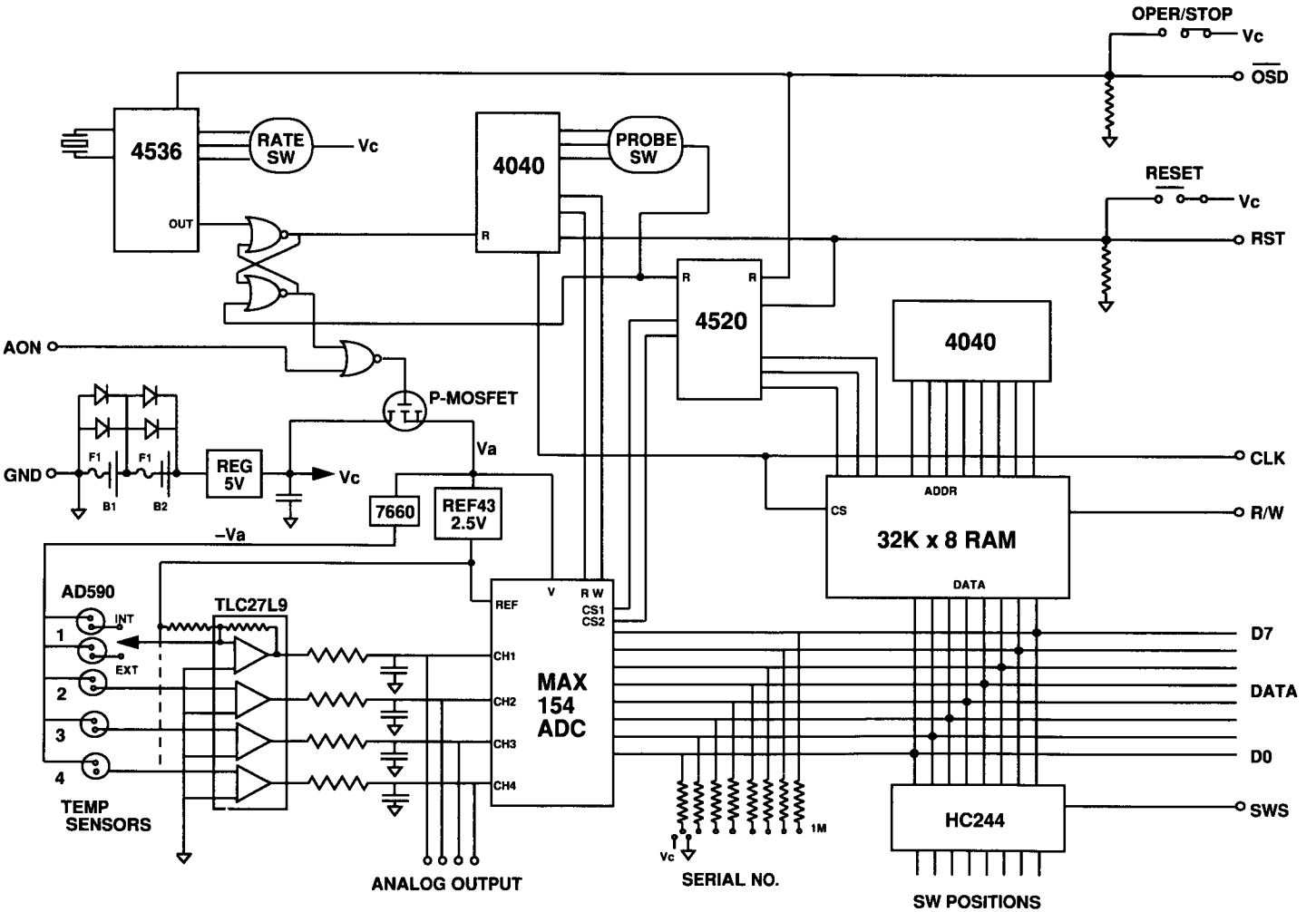


Figure 5. ATR-4 Circuit Schematic

N92-22722

FIBER-OPTIC PUSH-PULL SENSOR SYSTEMS

David L. Gardner, David A. Brown, and Steven L. Garrett
Code PH/Gd, Department of Physics
Naval Postgraduate School
Monterey, CA 93943

ABSTRACT

Fiber-optic push-pull sensors are those which exploit the intrinsically differential nature of an interferometer with concommitant benefits in common-mode rejection of undesired effects. Several fiber-optic accelerometer and hydrophone designs are described. Additionally, the recent development at the Naval Postgraduate School of a passive low-cost interferometric signal demodulator permits the development of economical fiber-optic sensor systems.

INTRODUCTION

Optical fiber sensors have been represented as being of rather remarkable sensitivity, even to the point of violating the Second Law of Thermodynamics [1], [2], [3], and easy to fabricate. Obviously, optical fiber sensor designs have not always realized these early expectations in sensitivity [4] and geometric flexibility. However, sensors using technology developed for the communications industry do offer the possibility of remote sensing, with no electrical power required for signal amplification, at distances of 20 km or more. Further, fiber sensors, and the compatible communication lines to them, offer several advantages, among which are the immunity from electromagnetic interference, light weight, low power requirements, low crosstalk, and reduced risk of shock or spark hazards [3]. All the sensors described in this paper exploit developments in communications technology; accordingly, each can be used to detect stimuli kilometers away from the transmitting light source and receiving electronics.

We divide optical fiber sensors into two general categories: intrinsic and extrinsic. Extrinsic fiber sensors are those in which the light, at some point in the sensing operation, leaves the fiber. An example is the Fabry-Perot interferometer, in which the light leaves the fiber and impinges on some sensing mechanism from which it is reflected. The reflected light is then combined interferometrically with the incident light, and the resulting optical fringe pattern is representative of the stimuli. Intrinsic fiber sensors are those in which the light never leaves the fiber; this type of sensor facilitates the relatively easy construction of push-pull optical fiber sensors.

In push-pull optical fiber sensors, an optical fiber coupler bifurcates the light power within a single fiber into two fibers forming the two legs of the interferometer. The two legs can then be made to operate in a differential mode, or push-pull, in contrast to sensors in which the sensor response to a stimuli is compared to an isolated reference leg. Since optical fibers are neither intrinsically sensitive or selective to stimuli [5], considerable difficulty exists in attempting to create an isolated reference insensitive to all environmental stimuli. Placing the two legs of a push-pull sensor in close proximity exploits the common-mode rejection intrinsic to an interferometric sensor, thus minimizing the effects of unwanted environmental effects. The sensor design effort can then focus on enhancing the stimuli of interest, and achieve greater sensitivity since both legs of the interferometer are responsive.

Interferometric fiber-optic sensors generate signals proportional to the phase differences between the light waves propagating in the two interferometer legs [6]. Since the signals are phase differences, the sensor is not hostage to fluctuations in the transmitted light power, a problem common to sensors whose output light intensity is proportional to the stimuli intensity.

Phase shift in an optical fiber can be caused by several mechanisms, such as variation in fiber length, polarization, index of refraction within the core, or variations in the core diameter. The optical phase shift, $\delta\phi$, is given in general form by [7]:

$$\delta\phi = \beta\Delta L + L\Delta\beta = \beta\Delta L + L\left(k\Delta n + \frac{\partial\beta}{\partial a}\Delta a\right) \qquad (1)$$

Here, β is the optical propagation constant along the fiber axis ($\beta=2\pi/\lambda$, where λ is the optical wavelength in the fiber core), L is the fiber length responsive to the stimuli inducing the phase shift, k is the optical wave number in vacuum, n is the index of refraction of the fiber core, and a is the fiber core diameter, typically 5-8 mm in the single-

mode fiber used for the sensors reported herein. Changes in any of the parameters in eq. (1) will induce a phase shift; for the sensors to be described, the length of the fiber is modulated by the stimuli of interest.

FIBER-OPTIC ACCELEROMETERS

Two designs for fiber-optic interferometric accelerometers have been developed at the Naval Postgraduate School (NPS). The first (Fig. 1) consists of a seismic mass supported by two mandrels around which are wrapped the two legs of the interferometer [8], [9]. The mandrels are constructed from castable soft rubber, having a Poisson's ratio of nearly one-half, and act as transformers of longitudinal compression into circumferential strain. When the accelerometer case is displaced, the seismic mass compresses one of the mandrels while relieving the other. Accordingly, the two fibers wrapped around the mandrels experience strains of opposite sign, a feature of push-pull sensor design. The fibers reject changes in ambient temperature and pressure by exploiting the common-mode rejection which is also a feature of push-pull sensors. The acceleration sensitivity of a sensor of this type, using a 640 gm seismic mass and eight meters of fiber in each leg, is 10,000 radians/g [10].

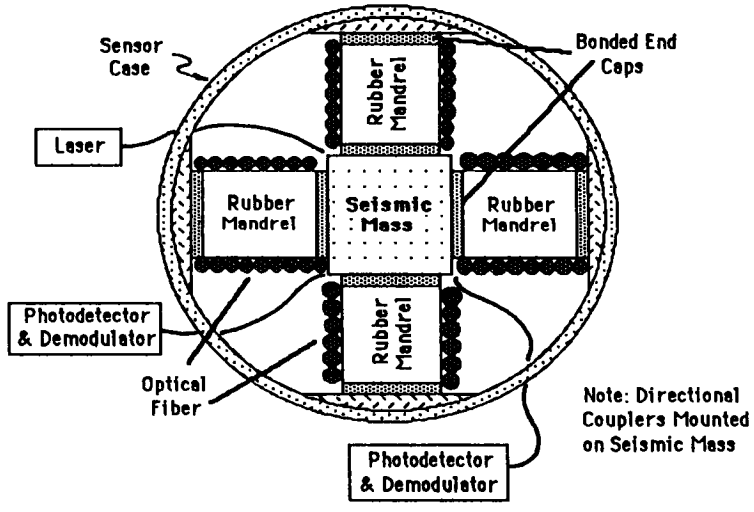


Figure 1. Sketch of a push-pull accelerometer using optical fiber wrapped rubber mandrels supporting a seismic mass. When the sensor case is displaced in a direction along the axis of a supporting mandrel pair, one of the mandrels is compressed while the other is relieved; accordingly, the optical fiber around one mandrel is stretched while the other is relieved.

Figure 2 is a plot of the interferometric signal output from this type of sensor compared to that of a high-sensitivity accelerometer. Novel techniques recently developed at NPS for demodulating the interferometric signals will be presented later in this paper.

Three such fiber-optic accelerometers placed on orthogonal axes and sharing a common seismic mass constitute a three-axis accelerometer. In all cases, the fiber-optic couplers used must remain within the accelerometer housing so as to not jeopardize the common-mode rejection of undesired environmental effects.

The second accelerometer design is sketched in Figure 3 [11], [12]. When the sensor is accelerated, strains are induced in the two thin circular plates. The disks are attached at the centers with the siesmic mass, which can be adjusted to achieve a desired resonance frequency. The two optical fiber legs forming the interferometer are wound in concentric, pancake-like coils, and are epoxied to each of the inside surfaces of the plates. When the sensor is displaced, the surfaces experience strains, of opposite sign, which are detected by the bonded fiber coils. As with the previous design, the two coils experience similar responses to unwanted influences of temperature and pressure by utilizing the common-mode rejection intrinsic to the interferometer. The acceleration sensitivity of a sensor of this type, using a 34 gm seismic mass and five meters of fiber in each leg, is 49 radians/g.

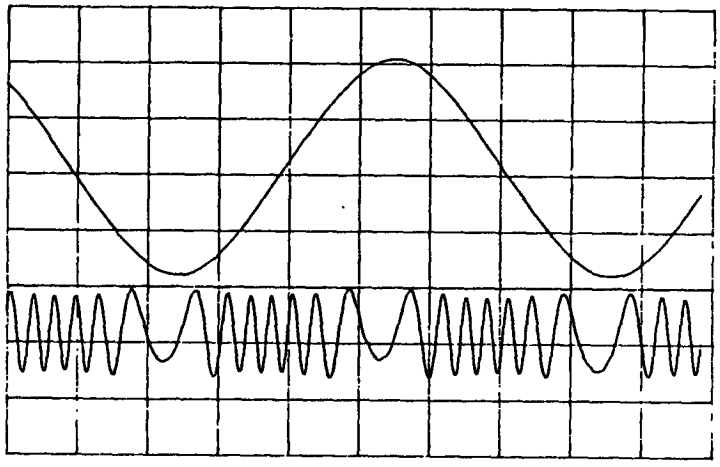


Figure 2. Sketch of an oscilloscope display of the output signal from a push-pull optical fiber sensor. The upper trace is from a conventional accelerometer and the lower trace is from an interferometric push-pull sensor.

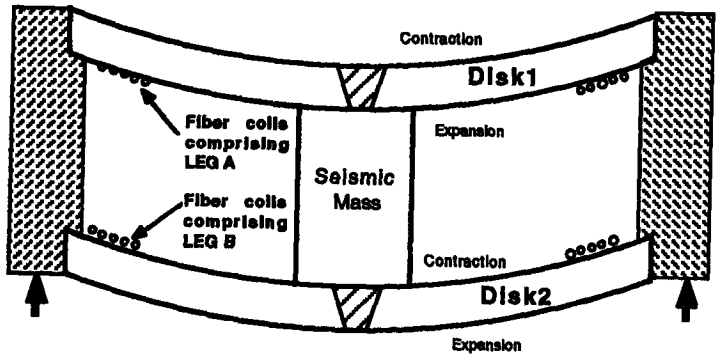


Figure 3. Sketch of a fiber-optic flexural disk accelerometer. When the case is accelerated, the inner surfaces of the two disks experience strains of oppossite sign. The optical fiber coils bonded to the surfaces also experience strains of opposite sign.

FIBER-OPTIC HYDROPHONES

Directional Hydrophones

If the above designs are placed within a neutrally-buoyant container and then placed in water [13], [14], the container will experience acoustically induced accelerations according to the linearized Euler equation:

$$\frac{\partial v}{\partial t} = \frac{1}{\rho} \nabla P, \tag{2}$$

where ρ is the fluid density. Accordingly, the sensor becomes bi-directional. The first sensor design is insensitive to changes in temperature or hydrostatic pressure (which deforms the neutrally buoyant case) since these effects cause identical changes in the two mandrels and are rejected by the differential behavior of the interferometer. The second design is insensitive since hydrostatic pressure causes both plates to experience similar strain.

Omni-directional Flexural Plate Hydrophone

The pressure-induced strain experienced by the two circular plates afforded the opportunity to construct an omni-directional sensor by placing the coils so as to enhance this response [15], [16], [17]. A sensor so constructed will be insensitive to acceleration, and the sensitivity can be increased by placing coils appropriately on both sides of the plates (Fig. 4). We have constructed and tested several of these "flexural plate" hydrophones. The theoretical sensitivity based upon the elastic properties of the plate materials and boundary conditions is in excellent agreement with the measured sensitivity for both acoustic and hydrostatic pressure in water and air.

We have also constructed four-coil versions of this dual-plate hydrophone made entirely from a high tensile strength castable elastomer having a low dynamic Young's modulus with a temperature coefficient of 0.3%/°C within the range of oceanic temperatures [18]. The four coils and coupler are cast into the unit and the gap between the plates is chosen so that the plates touch at the maximum operating depth. This permits the maximum sensitivity consistent with the breaking strain of the optical fibers since the hydrophone does not rely on the tensile strength of the plate material alone to support the hydrophone at crush depth.

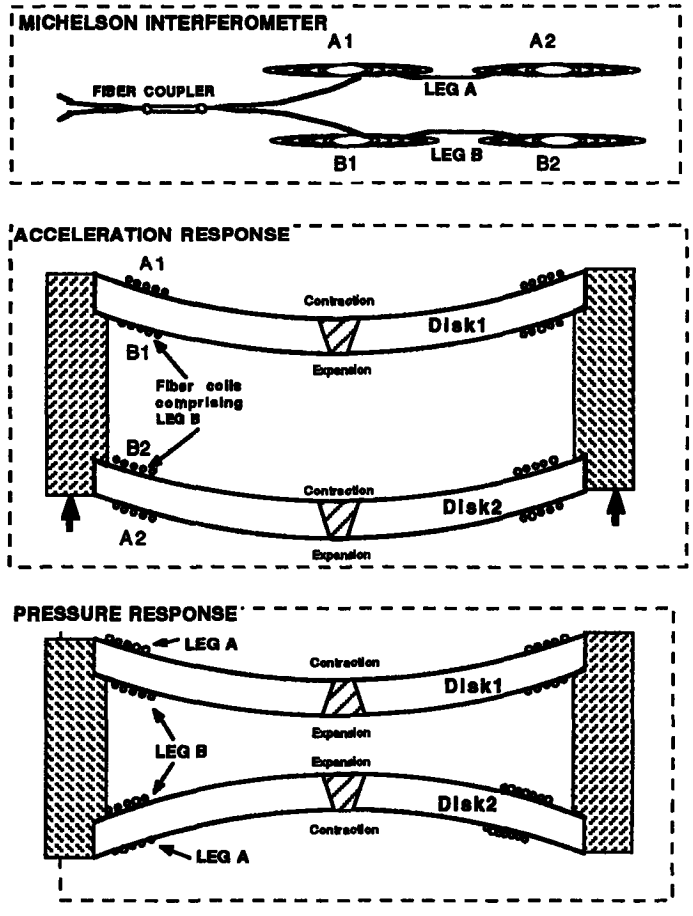


Figure 4. Sketch of a omni-directional flexural disk hydrophone interferometer and responses. The upper figure is a schematic representation of the optical fiber coupler and coils forming the Michelson interferometer. The middle figure illustrates the response of the hydrophone under acceleration; the two coils of either Leg A or Lag B experience no net change in length since one coil is stretched while the second is relieved. The lower figure illustrates the pressure response of the sensor; the strains experienced by both coils of Leg A and Leg B are of opposite sign.

Ellipsoidal Flextensional Hydrophone

When an oblate spheroidal shell, shown in Figure 5, having an aspect ratio $a/b > (2-\nu)^{1/2}$, where ν is Poisson's ratio, is subject to hydrostatic compression, the semi-major (a) axis and semi-minor axis (b) experience strains of opposite sign. If two optical fibers, which comprise the arms of an interferometer, are wound around the the semi-major axis and semi-minor axis circumferences of the oblate spheroid, pressure changes will induce a differential optical phase shift [19], [20].

As in the case of the flexural plate design, we have developed a "closed form" analytical description of this hydrophone which provides the necessary design parameters (a, b, and t) based on the choice of shell material, operating bandwidth, and maximum operating depth [21]. This design has the advantage of requiring no internal parts so that the shells can be fabricated by inexpensive methods (e.g. injection molded) after which the optical fibers can be attached to the exterior.

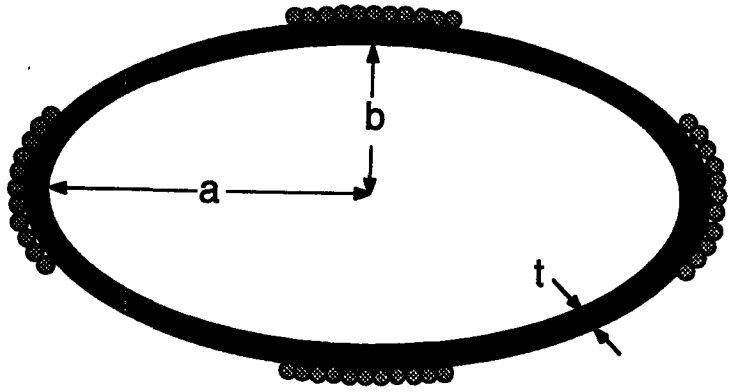


Figure 5. Cross-sectional sketch of a fiber-optic oblate spheroidal flextensional hydrophone showing the shell thickness, t, and the semi-major (a) axis and semi-minor axis (b) coil pairs. The three-dimensional shell is generated by rotation about the semi-minor axis (b).

This geometry also provides considerable flexibility in array design [22]. The hydrophone element or array to be "shaded" by varying the number of turns taken around each shell. Optical redundancy of the array to improve reliability can be accomplished by wrapping optical fiber of independent interferometers around the same shells. A pressure gradient hydrophone could be constructed by wrapping the fiber from one leg of the interferometer around the equator of one shell and the meridian of the other, and the opposite wiring for the other shell. A cardiod beam pattern could then be obtained by wrapping one leg of a second interferometer around the equators of both oblate spheroids and the other leg around the meridians so that the same two shells would give both the omni-directional and the bi-directional information. Clearly, higher-order arrays could be constructed in the same way.

Demodulation of Fiber-Optic Interferometric Signals

The most prevalent demodulation technique for 2x2 coupler-based interferometric fiber optic signals uses a laser modulated at a carrier frequency, while the phase shifts induced by the stimulus appear as side-bands [23], [24]. The modulation of the laser is accomplished by modulating the laser drive current, an approach introducing several problems among which are: 1) the requirement for laser diodes with long coherence lengths; 2) mode-hopping induced by the superimposed current modulation; 3) high setup complexity requiring FFT calibration of phase generated carrier spectral components to insure Bessel function balance and orthogonality; 4) the necessity of an optical path difference which increases susceptibility to additional laser phase noise and extraneous environmental effects; 5) fluctuations in sensor output due to scale factor instability, (scale factor quantifies the circuit's conversion of optical phase in radians to electrical output in volts), and 6) the signal induced by the stimulus appears as side-

bands of both the carrier frequency and at its harmonics; since not all harmonics are included, the total optical signal power is not available to correct for fluctuations in fringe visibility.

An ideal, equal split ratio, 3x3 coupler, has outputs with a relative phase difference of 120° [25]. This allows the generation of in-phase and quadrature signals without the real-time carrier waveform. Previous demodulator implementations using a 3x3 coupler used only two of the three available coupler outputs to produce the in-phase and quadrature signals, and these signals were again processed by the sine-cosine algorithm used for the phase generated carrier technique [26], [27]. Although this approach does alleviate many of the problems associated with phase generated carrier demodulation, the unavailability of stable 3x3 couplers postponed the practical implementation of this strategy. Recently, however, 3x3 couplers with good environmental stability and polarization insensitivity have become available [28].

In this demodulator implementation, all three outputs of the 3x3 coupler are used in a symmetric manner to re-create the phase modulating signal, $\phi(t)$ [25], [29]. No modulation of the laser wavelength is required, thereby reducing cost and complexity, and noise (there is less probability that the laser will "mode-hop"), while removing the previously described problems associated with laser modulation. The symmetric demodulation technique is capable not only of demodulating milliradian signals, it can also demodulate signals at the kiloradian levels, as opposed to the fractional radian upper limit of phase generated carrier techniques. Accordingly, the previously described push-pull sensors, which produce signals 20 to 50 dB greater than earlier sensors, can now be readily exploited. Further, this demodulator uses a simple feedback control circuit that is capable of maintaining a stable scale factor (volts/radian) in the presence of variations in both the total optical power and in the fringe visibility.

A Mach-Zehnder fiber optic interferometer consisting of a 2x2 coupler at the input and a 3x3 coupler at the output was fabricated in order to test the symmetric demodulation algorithm [Fig. 6]. The two legs of the interferometer were wrapped around separate piezoelectric cylinders which were driven 180° out-of-phase to produce large modulation amplitudes in the interferometer. The demodulator was implemented using low-cost electronic components such as AD712 and OP111 operational amplifiers, AD534 multipliers, and a DIV100 divider.

Typical phase generated carrier demodulation techniques have a one-to-one correspondence in the change in AC power and scale factor. This is very important, since, in principle, there is no way to distinguish fluctuations in scale factor from real signals. In this demodulator implementation, the scale factor is stable to better than ±5% even in the presence of optical power variations greater than 170%.

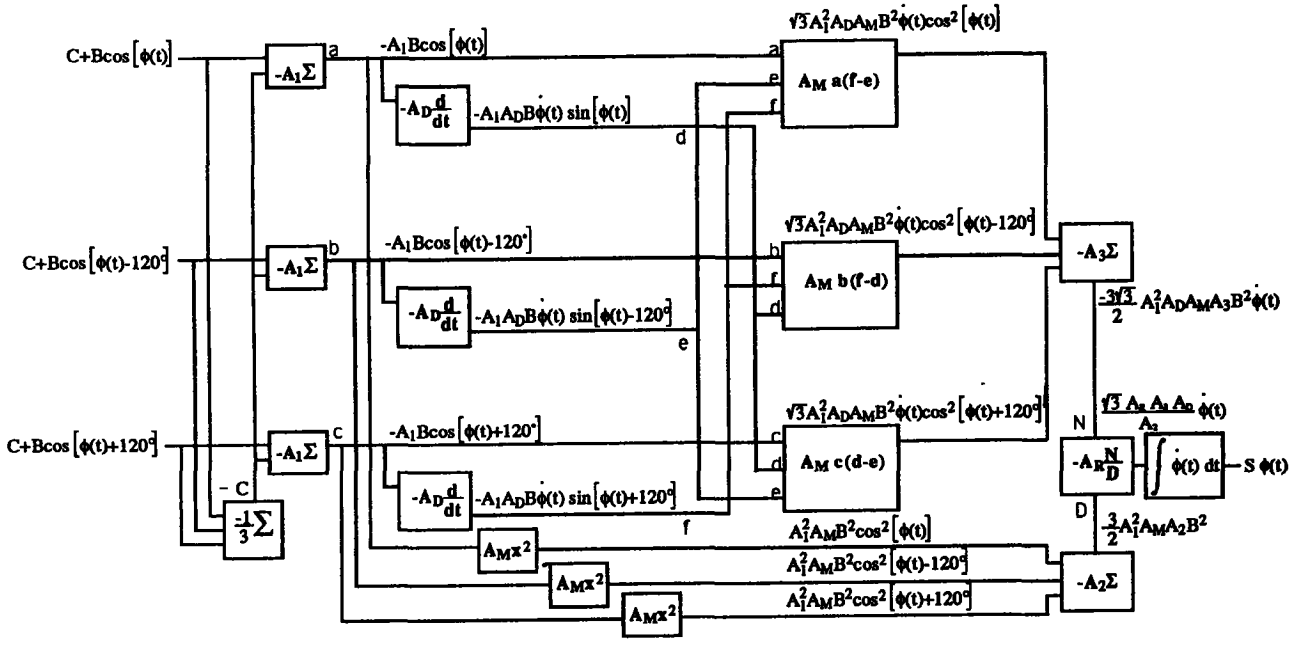


Figure 6. Block diagram of the passive symmetric interferometric signal demodulator signal flow using the outputs of a 3x3 optical fiber coupler. The outputs from each of the fibers are C + Bcos[ϕ(t)] and C + Bcos[ϕ(t) + 120°]. The stimuli-induced modulating signal, ϕ(t), is re-created at the output and is multiplied by scale factor, S.

The dynamic range of the demodulator is defined as the ratio of the maximum signal that can be accurately reproduced by the circuit to the minimum detectable signal in a given bandwidth. A conservative criteria defining the maximum signal was chosen as that amplitude for which the demodulated output produced a total harmonic distortion level (THD) of 4%. For the detectability, the bandwidth of the measurement has been normalized to 1 Hertz. The minimum detectable signal at 600 Hz is 220 μrad/√Hz and the 4% THD signal level is 140 radians, thus the dynamic range for this initial implementation is 116 dB [28], [29].

In evaluating the demodulator performance, the maximum and minimum signal levels were proportional to "1/f" (-20 dB/decade). For the maximum signal, this limit was due to the maximum fringe rate of the demodulator (65 krad/sec). The 1/f dependance in the minimum detectability was due to integration of the multiplier noise. This performance leads to a frequency independent dynamic range for the system, which is well suited for practical use, since laser noise, ocean noise, and many other systems exhibit noise power spectral densities which increases with decreasing frequency at low frequencies.

ACKNOWLEDGEMENTS

This work was supported by the Naval Postgraduate School Direct Funded Research Program, the National Oceanic and Atmospheric Administration, and the Naval Sea Systems Command.

REFERENCES

1. L. M. Lyamshev and Yu. Yu. Smirnov, "Fiber-optic sensors (review)," Akust. Zh. **29**, 289-308 (May-June, 1982).
2. T. G. Giallorenzi, J. A. Bucaro, A. Dandridge, G. H. Siegel, Jr., J. H. Cole, S. C. Rashleigh, and R. G. Priest, "Optical fiber sensor technology," IEEE J. Quantum Electron. **QE-18**(4), 626-665 (1982)
3. V. I. Busurin, A. S. Semenov, and N. P. Udalov, "Optical and fiber-optic sensors (review)," Kvantovaya Elektron. (Moscow) **12**, 901-944 (May 1985).
4. Thomas J. Hofler and Steven L. Garrett, "Thermal noise in a fiber optic sensor," J. Acoust. Soc. Am. **84** (2), 471-475 August 1988.
5. D. A. Brown, T. Hofler, and S. L. Garrett, " High snsitivity, fiber-optic, flexural disk hydrophone with reduced acceleration response," Fiber and Integrated Optics **8**(3), 169-191 (1989)
6. C. M. Crooker and S. L. Garrett, "Fringe rate demodulator for fiber optic interferometric sensors," in *Fiber Optic and Laser Sensors V*, Proc. Soc. Photo-Optical Inst. Eng. (SPIE) **838**, 329 (1987)
7. E. F. Carome, "Acousto-optic transduction in optical fibers and in fiber optic acoustic devices," in *Frontiers in Physical Acoustics*, edited by D. Sette (Soc. Italiana di Fisica, Bologna, 1986), pp. 476-505.
8. S. L. Garrett and D. L. Gardner, Multiple axis fiber optic interferometric seismic sensor, U. S. Patent No. 4,893,930, Jan. 16, 1990.
9. D. L. Gardner and S. L. Garrett, "Fiber optic seismic sensor," in *Fiber Optic and Laser Sensors V*, Proc. Soc. Photo-Optical Inst. Eng. (SPIE) **838**, 271 (1987)
10. A fiber-optic interferometric seismic sensor with hydrophone applications, LCDR D. L. Gardner, Doctor of Philosophy in Engineering Acoustics, September 1987. NTIS Report No. AD B112-487
11. D. A. Brown and S. L. Garrett, "An interferometric fiber optic accelerometer", in *Fiber Optic and Laser Sensors VIII*, Proc. Soc. Photo-optical Inst. Eng. (SPIE), **1367**, 282-288, (1990).
12. D. A. Brown, T. Hofler , and S. L. Garrett, Flex disc accelerometer, Patent application, Navy case 73054 (1991).
13. D. L. Gardner and S. L. Garrett, "High-sensitivity fiber-optic compact bidirectional hydrophone (U)," CONFIDENTIAL, U.S. Navy J. Underwater Acoust. JUA(USN) **38**(1), 1-21 (January, 1988)
14. D. L. Gardner, R. K. Yarber, E. F. Carome, and S. L. Garrett, "Fiber-optic interferometric geophone with hydrophone applications," 6th Int. Conf. on Integrated Optics and Optical Fiber Communications (Jan. 1987)
15. D. A. Brown, T. Hofler, and S. L. Garrett, "High-sensitivity, fiber-optic, flexural disk hydrophone with reduced acceleration response", Fiber and Intergrated Opt. **8**, 169-191 (1989).
16. D. A. Brown, T. Hofler, and S. L. Garrett, "A fiber-optic flexural disk microphone", in *Fiber Optic and Laser Sensors VIII*, Proc. Soc. Photo-optical Inst. Eng. (SPIE), **985**, 172-182, (1988).
17. T. Hofler and S. L. Garrett, Flexural disk fiber optic interferometric omnidirectional hydrophone, U. S. Patent No. 4,959,539, September 25, 1990.

18. S. L. Garrett, D. A. Brown, B. L. Beaton, K. Wetterskog, and J. Serocki, "A general purpose fiber-optic hydrophone made of castable epoxy", in *Fiber Optic and Laser Sensors VIII*, Proc. Soc. Photo-optical Inst. Eng. (SPIE), **1367**, 13-29, (1990).
19. S. L. Garrett and D. A. Danielson, Fiber optic interferometric ellipsoidal flextensional hydrophone, U. S. Patent No. 4,951,271, August 21, 1990.
20. D. A. Danielson and S. L. Garrett, "Fiber-optic ellipsoidal flextensional hydrophones", J. Lightwave Tech. 7(12), 1995-2002 (1989).
21. Optical fiber interferometric acoustic sensors using ellipsoidal shell transducers, D. A. Brown, Doctor of Philosophy in Engineering Acoustics, June 1991
22. S. L. Garrett and D. A. Brown, "Fiber-optic push-pull hydrophones", DoD Conference on Fiber Optics, McLean, VA, March, (1990).
23. A. Dandridge, A. B. Tveten, T. G. Giallorenzi, "Homodyne demodulation scheme for fiber optic sensors using phase generated carrier," IEEE Journal of Quantum Electronics, **QE-18**(10), October 1982.
24. A. Dandridge and A. B. Tveten, "Phase compensation in interferometric fiber-optic sensors," Optics Letters, 7(6), June 1982
25. C. B. Cameron, "Recovering Signals from Optical Fiber Interferometric Sensors", Ph. D. Dissertation, Electrical Engineering, Naval Postgraduate School, Monterey, CA, June, 1991.
26. S. K. Sheem, "Optical fiber interferometers with [3x3] directional couplers: Analysis," J. Appl. Phys. **52**, 960, June 1981.
27. K. P. Koo, A. B. Tveten, and A. Dandridge, "Passive stabilization scheme for fiber interferometers using (3x3) fiber directional couplers," Appl. Phys. Lett. **41**(7), October 1982.
28. M. A. Davis, A. D. Kersey, M. J. Marrone, and A. Dandridge, "Chartacterization of 3x3 fiber couplers for passive homodyne systems: Polarization and Temperature Sensitivity," paper WQ2, *Proc. Optical Fiber Communications Conference*, Houston, TX, Feb. 6-9, 1989.
29. C. B. Cameron, R. M. Keolian, and S. L. Garrett, "A Symmetric Analogue Demodulator for Optical Fiber Interferometric Sensors", Proc. 34th Midwest Symposium on Circuits and Systems (IEEE), Monterey, CA, May 14-17, 1991.

N92-22723

COMMERCIAL "CAPACIFLECTOR"

John M. Vranish
NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

A capacitive proximity/tactile sensor with unique performance capabilities ("Capaciflector" or capacitive reflector) is being developed by NASA/GSFC for use on robots and payloads in space in the interests of safety, efficiency, and ease of operation. Specifically, this sensor will permit robots and their attached payloads to avoid collisions in space with humans and other objects and to dock these payloads in a cluttered environment. The sensor is simple, robust, and inexpensive to manufacture with obvious and recognized commercial possibilities. Accordingly, NASA/GSFC, in conjunction with industry, is embarking on an effort to "spin" this technology off into the private sector. This effort includes prototypes aimed at commercial applications. The principles of operation of these prototypes are described along with hardware, software, modelling, and test results. The hardware description includes both the physical sensor in terms of a flexible printed circuit board and the electronic circuitry. The software description will include filtering and detection techniques. The modelling will involve finite element electric field analysis and will underline techniques used for design optimization.

INTRODUCTION

The objective for NASA purposes is to develop a proximity sensing skin that will permit a robot to sense intruding objects without blind spots (up to one foot). This is a muti-purpose sensor. When used as an array on its arms, the robot can be prevented from colliding with an object in space, particularly a human being. When sensing skin elements are placed on an Orbital Replacement Unit (ORU), these units can be manipulated, berthed and fastened down with unprecedented accuracy and safety-no possibility of unwanted collisions. NASA research has also demonstrated that scanning the sensor can produce clear images and that the near range resolution is so accurate that precontact virtual force control is possible. The sensor is capable of becoming central to NASA space robot control.

This sensing skin must be able to function reliably in the extreme environment of space and not disturb or be disturbed by neighboring NASA instruments. It should be simple, compact and be incidental to the robot design. An approach based on an array of capacitors appears promising in solving both the proximity and tactile models [1]. However, the system must be able to detect objects (including humans) at ranges in excess of one foot so that the robot can react. To obtain such a range, a capacitive sensor typically must be "stood off" from the grounded robot arm a considerable distance (approximately one inch). This would disfigure the robot arm, causing it to be bulkier than necessary. It would also make cross-talk between the sensor elements more pronounced and would likely impede the flow of heat from the robot arms to outer space (a serious problem for the Flight Telerobotic Servicer (FTS)). The "Capaciflector" (capacitive reflector) described in this paper solves these problems and, in so doing, advances the state-of-the-art in capacitive sensor performance.

NASA is now in the process of developing a commercial version of the "Capaciflector", 2.5 cm (1 in.) on a side. This sensor will be ultra simple, inexpensive and compact (essentially a piece of flexible printed circuit board with electronic circuitry mounted on its reverse side). But, with ranges in excess of 13 cm (5 in.) vs 2.5 cm (1 in.) for comparably-sized commercially available capacitive sensors, the commercial "Capaciflector" will use both an analog output (signifying range) and an on/off switching signal that can be reset as required.

THE "CAPICIFLECTOR"

The "Capaciflector" is a capacitive sensing element backed by a reflector element which is driven by the same voltage as the sensor to reflect all field lines away from the grounded robot arm, thus extending the range of the sensor. This approach is an extension of the technique used in instrumentation systems where a shield or guard is used to eliminate stray capacitance [2].

Fig. 1 shows the principles of operation in terms of charges and electric fields. Fig. 1a shows a capacitive sensor not using the "capaciflector" principle. Since we are using relatively low frequencies (approximately 20 kHz) we have the quasi-static condition and static charges and electric fields can be used to determine the capacitance the sensor "sees". We can see that the smaller the stand-off from the grounded robot arm, the larger the capacitive coupling between the sensor and ground. This, of course, has the effect of reducing the relative coupling between the sensor and the object being sensed and hence reducing sensor range and sensitivity. On the other hand, increasing the stand-off increases the bulk of the robot arm and adds wires and wiring complications. And, when the insulation materials are added to support the stand-off, the ability of the robot arm to dissipate thermal energy into space is impeded. When the "capaciflector" principle is used (Fig. 1b)[3], the field lines from the sensor are prevented from returning directly to ground. The effective stand-off is approximately the width of the active shield or capacitive reflector. Thus, we can have a skin with very little thickness (on the order of 0.060 inches) and a robot arm with very little bulk and still have the performance of a large stand-off. Fig. 2 shows the electronic circuitry. The capacitive coupling between the sensor and the object being sensed is used as the input capacitance tuning the oscillator frequency. As an object comes closer, the capacitance increases and the oscillator frequency decreases. On the other hand, the reflector is attached to the output of the voltage follower so it is electrically isolated and prevented from affecting the tuning of the oscillator frequency. At the same time, the voltage of the reflector follows that of the oscillator. Thus, the reflector is in phase with (and reflects) the electric field of the sensor without being affected by the coupling between the sensor and an approaching object.

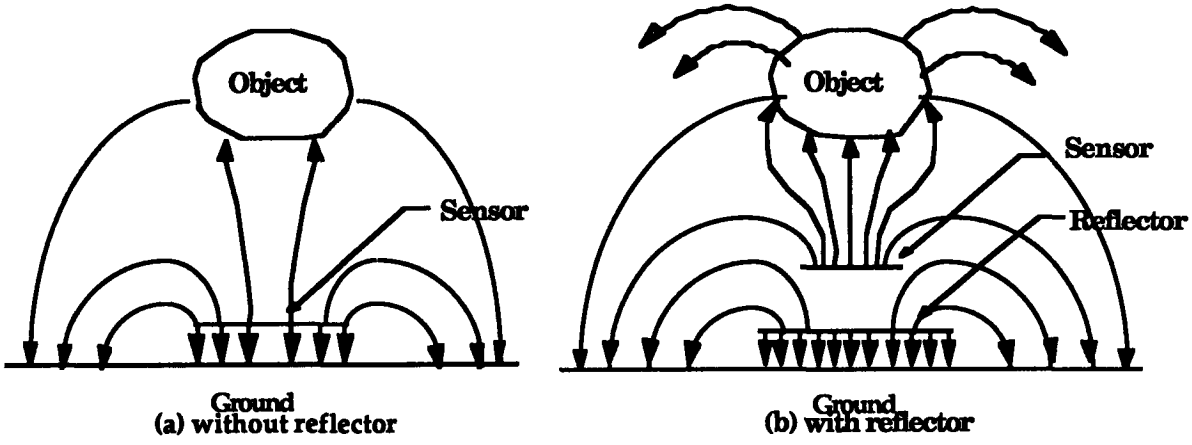


Fig. 1 : "Capaciflector"principle[3]

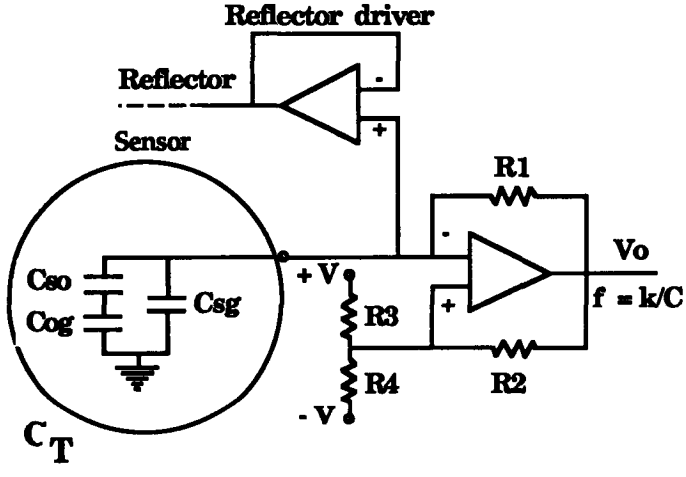


Fig. 2 : "Capaciflector"circuitry[3]

DETECTION

We will now examine the means by which the sensor detects an object[3]. The discussion will be limited to conductors for simplicity although dielectrics are also easily detected. Both the grounded and ungrounded (Fig.1) cases will be examined. Since we have low frequency, (approximately 20 kHz), the quasi-static case holds. Assuming a momentary positive potential V in Fig. 1b. we can see that the electric field lines emanating from the sensor towards the object induce negative charges on the object surface nearest the sensor. Thus that surface can be considered one plate of a capacitor and the sensor the other. But, an ungrounded conductive object is charge neutral so an equal amount of positive charge will form on the surface away from the sensor so as to ensure that there is no net electric field in the conductor. These charges couple back to ground which creates a second capacitor in series with the one mentioned above. These are labeled in Fig. 2 respectively as:

$$C_{so} \text{ and } C_{og}$$

But; there also is a path where the electric fields from the sensor can go around the active shield and couple to ground directly. This is labeled as:

$$C_{sg}$$

Thus our tuning capacitance is:

$$\frac{C_{so} C_{og}}{C_{so}+C_{og}} + C_{sg} = C_t \qquad (1)$$

In the case where the object is grounded, equation (1) reduces to:

$$C_t = C_{sg} + C_{so} \qquad (2)$$

Examining equations (1) and (2) above, since we are looking for small changes in C_t it is clear we want C_{sg} to be small. Therefore, we want the shield or reflector to force the field lines from the sensor towards the object as much as possible.

We now turn to the case where the object is not grounded [4,5,6]. We know:

$$C = \frac{Q}{V} \qquad (4)$$

We also know that a good conductor must have the same potential everywhere on its surface. Therefore the potential on the object will be that of its furthest point from the sensor. we will call the potential on the sensor V and the object potential V_o. Thus we have:

$$\frac{Q_i}{V - V_o} = C_{so} \text{ and } \frac{Q_i}{V_o} = C_{og} \qquad (5)$$

where

Q_i = charge induced in the object.

It is apparent that an object with any dimension more than a few inches in any direction (for example length) forces the potential on the entire surface of the object to be very low. And, as the experimental evidence shows,in practice, all objects are approximately grounded.

Electric field details

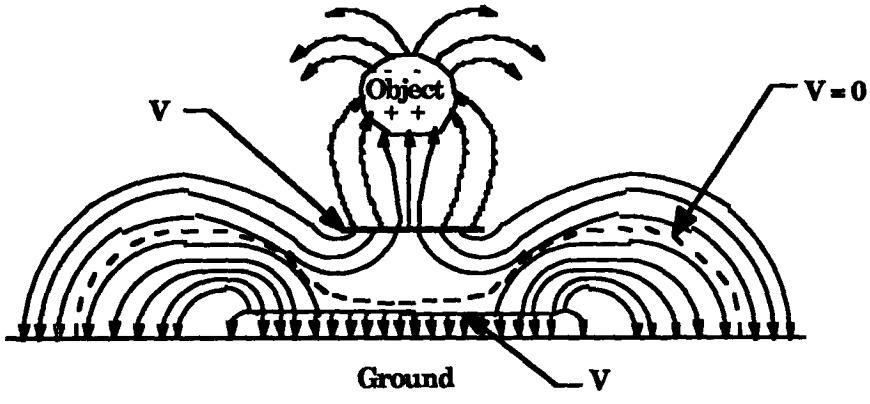


Fig. 3 : Electric field details[3]

MODELING

In order to verify the experimental results, and to further improve the sensor, a static electric field model was developed. The objective was to determine the percentage change in frequency of the oscillator resulting from the introduction of an object within the field of the capaciflector. The frequency of oscillation of the circuit in Figure 2 can be shown to be

$$f = \frac{\ln 0.5}{2R_1C}$$

Where: $R_3 = R_4 = 2R_2$

This implies

$$\frac{\Delta f}{f_0} = \frac{\Delta C_t}{C_{t0} + \Delta C_t}$$

where f_0 and C_{t0} represent the frequency and the capacitance of the sensor in the absence of an object, and Δf and ΔC_t represent the change in frequency and capacitance respectively because of the introduction of an object.

The method of moments approach was chosen to determine the capacitances, as it allows one to model systems with no boundaries. The modeling approach is therefore similar to that used by Volakis et al. [7]. In our case, the system consists of the grounded robot arm, the sensor shield, the sensor, and the detected object. The system is approximated by a two dimensional model — we solve the problem for a cross section of the system assuming that the system extends to infinity along the axis perpendicular to the plane.

In this method, after discretizing the two dimensional system entities, each discrete element of length Ds with a charge density of r, is approximated to a point charge of magnitude r Ds, located at the center of mass of the element. The charge densities are then determined by solving the following set of M linear equations, M being the total number of elements in the system.

$$\sum_{m=1}^{M} \rho_m K_{nm} = V_n - V_k ; n = 1,2,\ldots,M \ , n \neq k$$

$$\sum_{m=1}^{M} \Delta s_m \rho_m = 0$$

where V is the voltage of an element, K_{nm} is the integral of the two dimensional Green's function for statics, and k reference element. K_{nm} is given by the following relations [Ibid]:

$$K_{nm} = \Delta s_m \ln(r_{nk} / r_{nm}) / 2\pi\varepsilon ; m \neq n , n \neq k$$

$$K_{nm} = \Delta s_m \left[\ln\left(r_{nk} / \Delta s_m \right) + \Delta s_m (1 + \ln 2) \right] / 2\pi\varepsilon ; m = n$$

where e is the permittivity of the medium and r_{ij} is the distance between the ith and the jth elements (point charges).

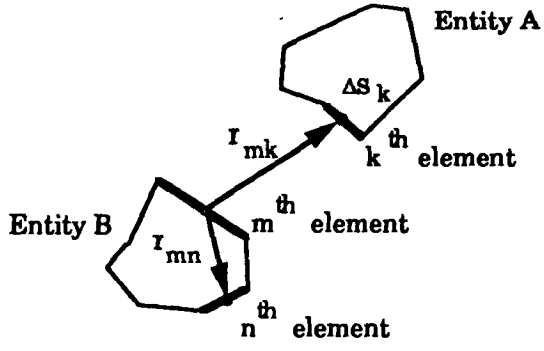


Fig. 4: Description of the variables used in the method of moments

The sensor capacitance is then determined by summing the capacitances of the elements representing the sensor:

$$C_s = \sum_{m=1}^{N} \Delta s_{s_m} \rho_{s_m} / \left(V_{s_m} - V_k \right) ; \rho_{s_m} > 0$$

where N is the total number of elements representing the sensor.

A computer program was written to calculate the total capacitance seen by the sensor for several configurations. For each configuration, the program moves a circular object on a grid placed on the plane of the entities, and the sensor capacitance is computed for each object position. One of the outputs the program provides is a data file for drawing the frequency change vs distance plots for each such configuration.

The program was used to plot graphs for the four configurations shown in Fig. 5 which were tested in the laboratory, and the results are shown in Fig. 6. The abscissa represents the distance of the object from the sensor, and the ordinate the percentage change in frequency. The center of the object is above the center of the sensor for the

cases shown.

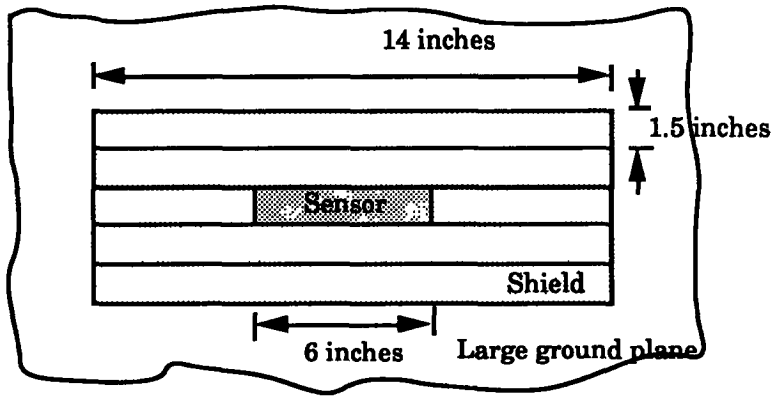


Fig. 5a : Test sensor

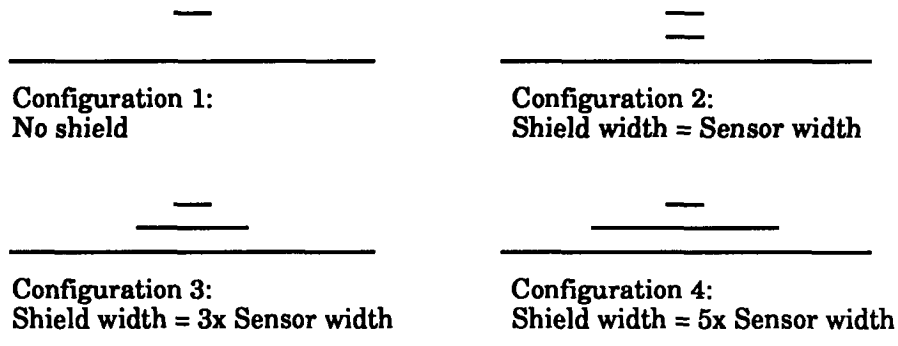


Fig. 5b: Sensor/Shield configuration

Fig. 5: Test configurations [3]

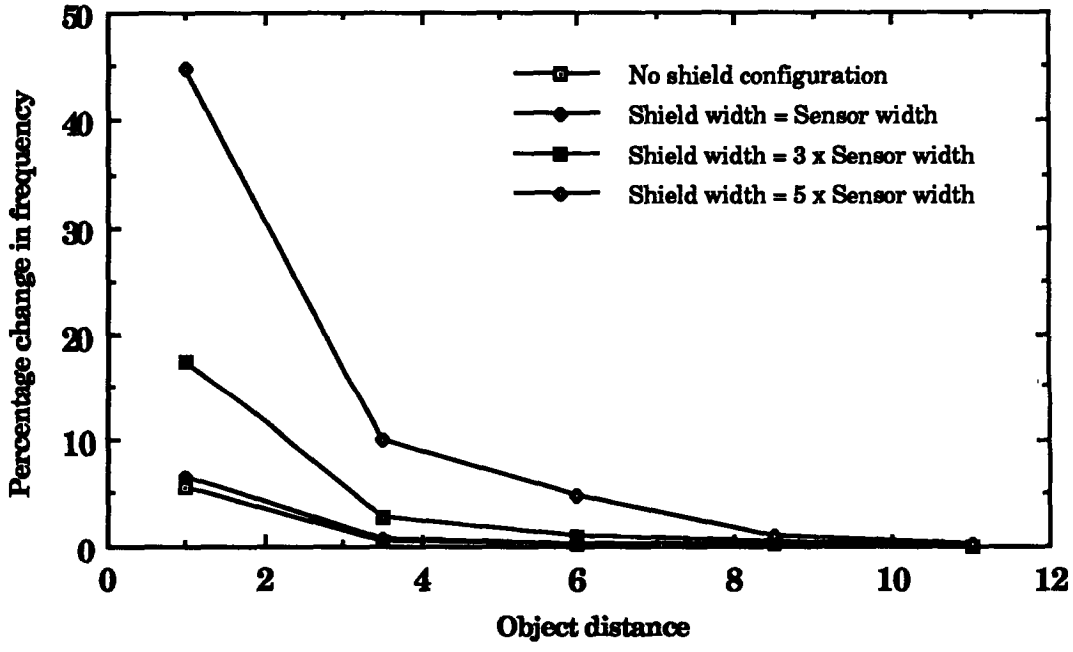


Fig. Modeling results, [3]

% Frequency Change vs Object Distance From Sensor.

EXPERIMENTAL RESULTS

An experimental laboratory set-up was assembled and a similar set of sensor configurations and object positions measurements were taken. The results of the experiment are shown in Fig. 7 and are similar to the model results. Since the computer model only simulates a two dimensional configuration, the results of the simulation assume infinitely long strips of the sensor, reflector, and object. The experimental set-ups were similar; the sensors was approximately six inches long, the reflector approximately fourteen inches long, and the object one inch in diameter and thirty six inches long. The reflector was made from strips of copper foil that could be connected in the configurations shown in Figure 5b. Subsequent testing has shown that the sensor must be shorter than the reflector to reduce end effects which substantially reduce sensitivity. The explanation is that the reflector must totally surround the sensor to contain the field. Otherwise, the flux lines from the sensor will simply shift to the lower field strength and return to the ground at the ends of the sensor, thereby reducing the coupling to the object.

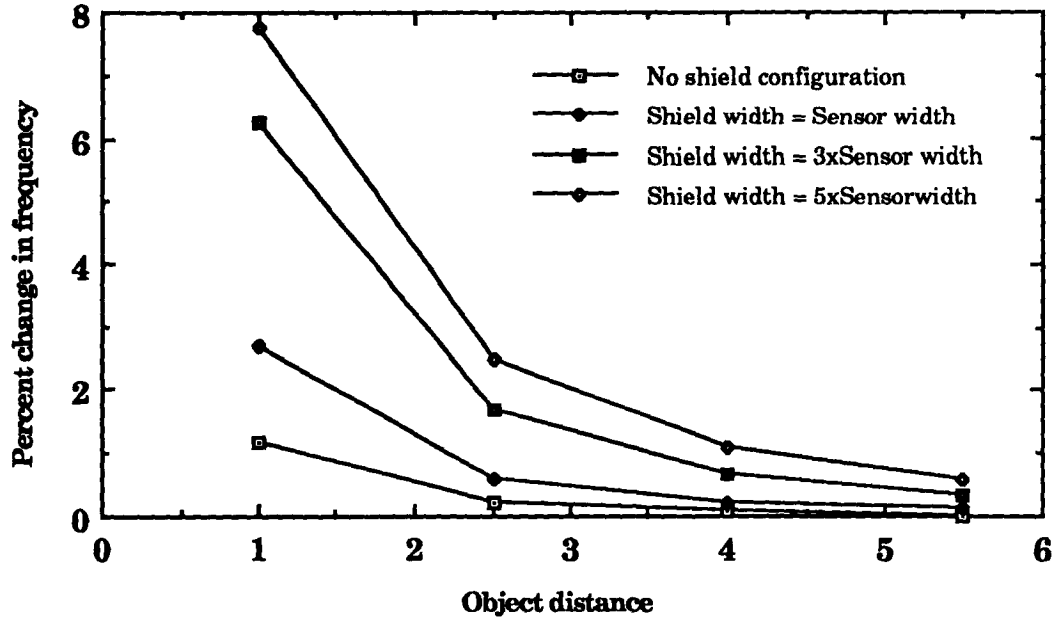


Fig. 7

Experimental results [3]

DISCUSSION OF RESULTS

The results from the modeling and the experiment are similar. Both show the frequency change is inversely proportional to the object distance distance from the sensor. They both show that the sensitivity increases dramatically as the the shield width increases. The increase is approximately 7-fold for the experimental result and almost 9-fold for the model.

The substantial difference shown between the modeled results and the experimental results are probably due to our primitive models used to date. The model program assumes infinitely long strips for the sensor, shield, and object, while our experiment used a 6 inch sensor with 14 inch shield. End effects or the short sensor may account for the difference; our modeling has not progressed far enough to determine. The rate of variation between the curves is also different. The model shows almost no difference between the curves for no shield and shield=sensor width, while the experimental results show a substantial difference. This result may be entirely due to inaccuracies in the model. Similarly, there is a difference between the rate of change between the upper two curves on the graphs. The model shows an increasing rate of change difference while the experimental result shows almost a constant difference. We cannot presently account for this result, but it may be due to either the model or to electronic circuit limitations. This latter conjecture comes from the fact that the frequency changes are substantial and nonlinearities may limit the frequency shift. Investigations are continuing.

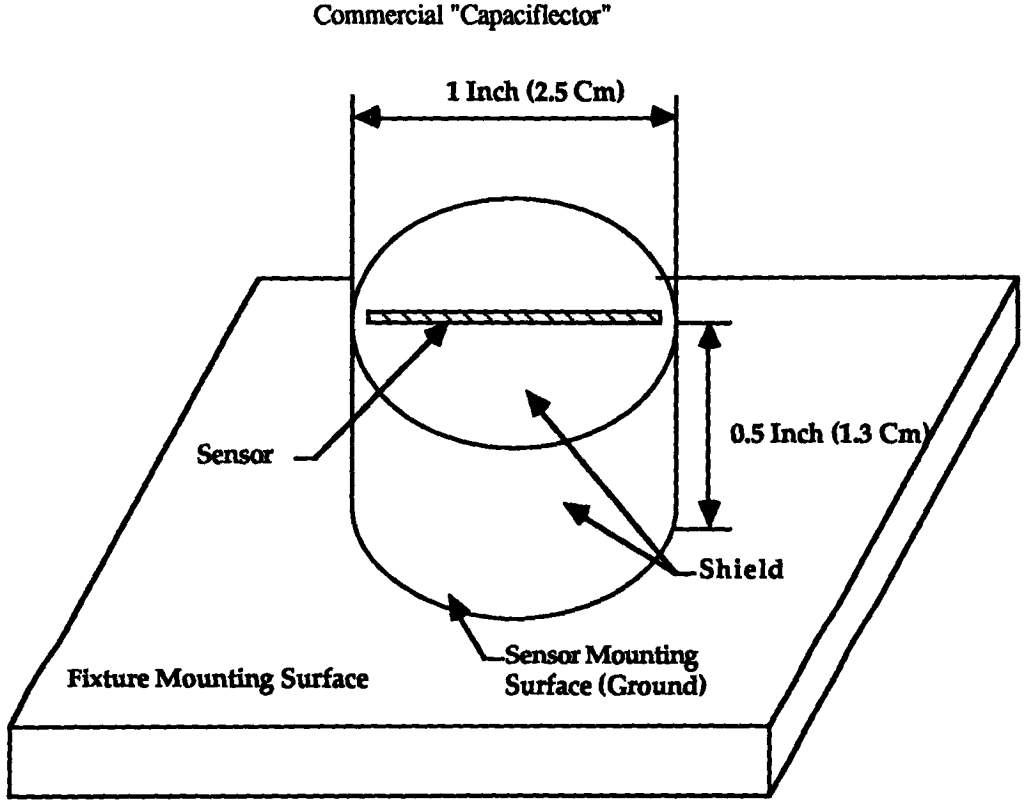


Fig. 8: Commercial "Capaciflector"

The commercial "Capaciflector" (Fig. 8) should be as small as possible so that it can be placed in grounded, confined areas, the typical situation in industrial applications. Thus, its reflective shield will be smaller than that of the NASA space version and, accordingly, the range and sensitivity/dynamic range will be compromised. For a sensor 1 in. dia and 0.50 in thick, we have measured a range of 7 in. ,when mounted on an insulator, 5 in. when mounted on a conductor. This is still 5 to 7 times the range and sensitivity achieved by commercially available capacitive sensors in a package 1/2 the volume (to include the sensor head and electronics-signal amplification and filtering). Thus, we have a significant increase in performance. In addition, the cost of fabrication will be much reduced. The sensor head and input/output leads is essentially a flexible printed circuit board. The electronics is ultra simple and straight forward and will be attached to the reverse side of the sensor.

With such an increase in performance, it seems sensible to have an analog signal which measures range in addition to the customary on/off switching output. Furthermore, the switching output will be electronically adjustable so that the sensor can be calibrated at the worksite, in real time, to respond to any of several object shapes and materials at any of several ranges.
These features will permit the sensor to have unprecedented performance and flexibility. It will be able to be used in the classic sense as a switching noncontact sensor. And, in this role, it will have a marked advantage over existing sensors with a greater range and superior signal to noise ratio in that range. Or alternately, it can provide a crisper, more certain switching point at a less than maximum range to include a crisper detection of edges. And, with the

electronically adjustable switching output, it will be able to be calibrated against a given object of a given material, at a given range in real time. This will result in extraordinary operational flexibility. as well. It can be used to determine range, which is also unprecedented for industrial-type capacitive (or inductive) sensors. And, with the superior range available, close-in range determination now becomes practical. This, in turn, suggests an entire new dimension in adaptive control for automated techniques in robots and machine tools. For example, automated, machine contour following will become practical. Also, robots and machine tools will be able to slow down just before contacting an object, switch from position control to force control mode before contact and thereby minimize the shock inherent in contacting the object and the instabilities that attend. This will also minimize the need for passive compliance; add-on solutions that have been used in the past.

SUMMARY

The NASA "Capaciflector" is well on its way towards becoming a central part of collision avoidance, docking and berthing and pre-contact force control in space and will be used extensively on robot arms, robot end effectors and payloads. It will also likely find uses inside robot mechanisms in support of their controllers. Thus GSFC has a large and growing in-house capability regards this sensor and this capability can easily be applied to the Commercial "Capaciflector".

REFERENCES

1. Vranish, John M. and Durg S. Chauhan,*"Tri-Mode Collision Avoidance Skin for Robot Arms in Space"*, Proc. Third International Symposium on Robotics and Manufacturing, 1990.

2. Webster, John G. , *"Tactile Sensors for Robotics and Medicine"*, John Wiley & Sons,Inc., Copyright 1988, pp. 209,202.

3. Vranish, John. M., McConnell, R. L., Mahalingam, S., *The International Journal of Computers and Electrical Engineering, ["Capaciflector" Collision Avoidance Sensors for Robots]* to be published.

4. Hayt, William H. Jr.,*"Engineering Electromagnetics"* , McGraw-Hill, Inc., Copyright 1981, 1974, 1967, 1958, p.146.

5. Lorrain, Paul, Corson, Dale R. and Lorrain, Francois, *"Electromagnetic Fields and Waves"*, W.H. Freeman and Co.,Copyright 1988,pp.108-116.

6. Fischer, C. R. *"Build the Digital Theramin"*, 1989 PE HOBBYISTS HANDBOOK, pp. 142,143,144,152.

7. Volakis, J.L., et al. *"Mapping of electrostatic fields using the IBM personal computer"*,IEEE Transactions on Education, Vol E-30, No. 4, Nov 1987, pp 247-250.

ENVIRONMENTAL TECHNOLOGY

(Session E4/Room C1)

Thursday December 5, 1991

- **Water Quality Monitor**
- **Remote Semi-Continuous Flowrate Logging Seepage Meter**
- **Calcification Prevention Tablets**
- **Automated Carbon Dioxide Cleaning System**

N92-22724

WATER QUALITY MONITOR (EMPAX INSTRUMENT)

Warren C. Kelliher
NASA Langley Research Center
Hampton, VA 23665

Ben Clark and Mike Thornton
Martin Marietta Astronautics Group
Denver, CO 80201

ABSTRACT

The impetus of the Viking Mission to Mars led to the first miniaturization of a X-ray Fluorescence Spectrometer (XRFS). Two units were flown on the Viking Mission and successfully operated for two years analyzing the elemental composition of the Martian soil. Under a Bureau of Mines/NASA Technology Utilization project, this XRFS design was utilized to produce a battery powered, portable unit for elemental analysis of geological samples. This paper will detail design improvements and additional sampling capabilities that have been incorporated into a second generation portable XRFS that was funded by EPA/NASA Technology Utilization project. The unit, known as EMPAX (Environment Monitoring with Portable Analysis by X-ray) was developed specifically for quantitative determination of the need of EPA and any industry affected by environmental concerns, the EMPAX fulfills a critical need to provide on-site, real-time analysis of toxic metal contamination. A patent has been issued on EMPAX, but a commercial manufacturer is still being sought.

INTRODUCTION AND BACKGROUND

In 1976, two miniaturized X-Ray Fluorescence Spectrometers (XRFS) were landed on Mars and successfully operated for over two years performing elemental analysis of Martian soil samples. These spectrometers were built for NASA by Martin Marietta and were part of the instrumented landers of the Viking Mission to Mars Project. These spectrometers used proportional counters for X-ray detectors, radioactive isotope sources as the X-ray generators and standard pulse height analysis for the electronics. This combination has served as the basis of all national and international space missions flown to date.

In 1979, NASA and Martin Marietta built a portable battery operated version of this instrument and delivered this prototype instrument to the Bureau of Mines for evaluation. Since then, several bench and portable versions are commercially available with the proportional counters/isotope source combination.

In 1979, at the request of EPA, NASA and Martin Marietta embarked on a feasibility study to determine if the portable X-Ray Spectrometer concept could be advanced to the point where sensitivities to a series of toxic metals in the range of parts per million (ppm) to parts per billion (ppb) in water samples could be achieved. Through the use of precipitating agents and collection of the precipitate on filtration membranes, sufficient preconcentration was obtained to achieve the majority of the elemental sensitivities desired using an X-Ray tube as the X-ray generator and a solid state, liquid nitrogen cooled Si(Li) detector as the X-ray sensor. Based on this information, funding was obtained to develop and deliver two prototype instruments for evaluation by NASA and EPA.

INSTRUMENT DEVELOPMENT AND DESIGN

In the development of a portable XRFS for the EPA, two decisions were made early in the program that directly affected the final instrument design. One was to have separate modules for the basic components rather than have them all incorporated into one combined instrument package and two was to use commercially available components wherever possible. Since the instrument concept evolved in the 80's time period, it will reflect the technology that was available during that time. With some of the 90's technology, selected components could be changed or updated and this will be discussed in a later portion of this presentation. The instrument design was also affected by the changing environmental concerns and issues at EPA in the 80's. The changing emphasis on water quality and soil analysis capabilities varied sufficiently, that the instrument is equally adapt in its capacity to handle either type of sample.

PRECEDING PAGE BLANK NOT FILMED

The unit as presently configured consists of four modules: 1) Water Sample Module, 2) Portable Analyzer Module, 3) Electronics Module and 4) Computer Module.

Water Sample Module

The water sample module is a unit designed by Martin Marietta to accept known volume water samples that have had a precipitant (Dibenzyldithiocarbomate) and a buffer (Potassium Hydrogen Phthalate) added and will perform the filtration process by depositing the precipitate on a membrane filter that is housed for ease of handling in a 35mm slide mount holder. The analyzer module will fit onto the water module and once the filtration step is complete (as indicated by filtration complete light), the operator has only to slide the 35mm holder from the filtration position to the analyzer position to initiate sample analysis.

Portable Analyzer Module

The portable analyzer module or sensor head is a combination of commercially procured instruments that were modified and incorporated in an enclosure designed by Martin Marietta. The detector is a Si(Li) solid state detector cooled by liquid nitrogen from a reservoir (dewar) that has sufficient capacity for 8-12 hours of operation. When not being operated in the portable mode, the dewar is attached by hose to a larger liquid nitrogen dewar and is automatically refilled when the liquid nitrogen is depleted. This detector and dewar system was procured from PGT, Inc. Due to the high elemental sensitivities needed by EPA, it was deemed necessary to go to the high resolution solid state detector rather than use the proportional counter tube type detector. Likewise, it was necessary to eliminate the use of radioactive isotopes as the X-ray source and go to a miniaturized X-ray tube for both the safety aspects as well as the need for the high intensity fluxes obtainable from a tube source. A 30 Kilovolt, molybdenum target, X-ray tube was built for this unit by X-Tech, Inc. and included as a part of the analyzer module. The high voltage bias and electronics power is supplied from an array of lead-zinc "Gel-cell" batteries located in the back section of the analyzer module.

Electronics Module

The electronics module is the pulse height analyzer, used to accumulate the energy pulses from the detector into an energy versus counts spectrum. Rather than attempting an extensive modification of the electronics used in the portable unit for the Bureau of Mines, a commercially available unit, the Canberra Series 10 Multi-channel Analyzer, was chosen instead which required only minor modification to interface with the analyzer and computer modules. It is capable of storing up to 16 spectra before transfer to computer storage is required. The rechargeable NiCad batteries power the unit and also provide detector bias and detector electronics power.

Computer Module

The computer module is used to accept and store the spectra received from the electronics module and to convert the spectral peaks to quantitative concentration values. The 80's saw considerable changes and increasing capabilities in small portable computers and as a result our original choice of the only mini-computer available at that time, the HP-85, rapidly became obsolete and so for our field tests, we switched to a Grid Case 2 portable unit. This too is rapidly on its way to being replaced with more powerful units, but the point is that our software as well as comparable commercial versions, will all work satisfactorily on 286 or higher number CPU machines.

The combination of the four modules comprise what we have designated as the "EMPAX" instrument (EMPAX being the acronym for Environmental Monitoring with Portable Analysis by X-Ray). The EMPAX instrument has been issued Patent No. 5014287.

FIELD DEMONSTRATION TEST

Two prototype units were built by Martin Marietta, one of which was delivered to EPA, Las Vegas Laboratories in Nevada and the other to NASA, Langley Research Center in Virginia. Before delivery, both units were involved in extensive field testing. To illustrate the potential usage of these units, the field trip to a Superfund site in Aspen, CO, conducted in August 1988 will be described. The site was chosen for being in driving distance to the Martin Marietta facilities in Denver, CO, and for having been already characterized for toxic pollutants which

were chiefly mine tailings from operation of the mines in Smuggler Mountain. Working from dawn to dust at the site, the combined team of NASA, Martin Marietta and EPA personnel collected over 120 spectra at which time all the instruments and personnel needed regeneration. A normal survey of a site such as this one, involves the analysis of many samples in a certified laboratory, would take many months to obtain analysis results and would cost up to $100,000. The EMPAX instrument, in a few days, could also perform the site survey, would provide in-site analysis results and the total instrument fabrication cost is estimated at about $80,000. Thus, the instrument is very cost effective compared to standard analysis costs.

Water samples were also available at this site and were taken from the nearby outcrop stream. The analysis results surprised us with very high Z peaks that turned out to be Uranium. Since the water percolated through the abandoned and flooded old mine shafts, by hindsight, the presence of Uranium is not too unusual but demonstrates that XRFS because of its elemental scanning ability, will routinely detect the presence of unusual elements. Other analysis techniques such as Atomic Absorption or Emission Spectroscopy (ICP) require the preselection of elements of interest and this would miss the presence of elements that may also be of interest.

Comparison of the XRFS analysis to that performed by standard laboratory procedures as stated in the EPA evaluation report - "Precision and accuracy of the in-site measurements were within ±10% of the true value when compared to the samples analyzed in the laboratory." These results are considered outstanding by all parties concerned and exceed expectations for the analytical capabilities of the EMPAX in most situations.

POTENTIAL DESIGN CHANGES

The prototype instrument design was preset several years ago and with the continuing advances in the field of XRFS instrumentation, there are a number of changes that could be made to EMPAX to enhance its capabilities and commercial applications.

Electronic Module

In many of the commercial laboratory XRFS, the function of the PHA has been incorporated on a computer circuit board so as to use the computer terminal and keyboard functions as the PHA readout and controls. With the newer laptop computers being able to add full size boards, it is quite possible to combine the functions of the Canberra Series 10 Multichannel Analyzer into the Grid Case type computer and thereby eliminate one module entirely. The power supply functions of the PHA could then be transferred to the analyzer module.

Water Sample Module

The present design uses a membrane filter to collect the precipitate and thus achieve up to a 100 to 1 increase in concentration of the elements to be analyzed. With the advent of new ion exchange type membranes, it is possible to have the membrane filter act as the concentrator thus eliminating a time consuming filtration step. Since the ion exchange membrane can also act as a filter, it can also serve to collect the precipitate while concentrating the cations that do not react with the precipitant. By analyzing both the front and backside of the membrane filter, one can increase the number of cations that can be analyzed.

Analyzer Module

When initially designed, the portable liquid nitrogen cooled detector was a major advance over the use of proportional counter tubes. Now there are several commercially available units but the continual need of liquid nitrogen as a coolant is an impediment that could be rectified by changing to a thermoelectrically cooled detector. There are commercial available XRFS systems that use this means of cooling but power consumption is in order of 30 to 80 watts. The literature and experimental work by Martin Marietta has shown this power consumption can be reduced to 12 watts or lower which is within the ability of using batteries as the power supply. This change would eliminate the problem of obtaining or carrying liquid nitrogen to remote sites.

The prototype instrument lack of sensitivity to Cd, Sb and Ag is due to the choice of using Mo as the X-ray tube target source. The L line energies of these elements are obscured by the Argon present in the atmosphere and the K lines energies are too high to be excited by the Mo X-ray source. A possible simple means to enhance the sensitivity to these elements is to use some of the gaseous nitrogen from the boil off of the liquid nitrogen coolant

and have it act as a purge over the sample to be analyzed. A more effective but more complex solution is to be able to selectively change the excitation source energy. As part of this project, a higher voltage (50 Kv) miniaturized X-ray tube was devised that provided dual energy outputs through the use of either a composite single target or a dual target source. This development came too late to include in the prototype instruments but have since become commercially available.

COMMERCIALIZATION

The prototype EMPAX instruments were built mainly to demonstrate feasibility and be used for field evaluation studies. During the design phase, consideration was made to make fabrication as simple as possible by using commercially available components, however modification and redesign is still needed if EMPAX is to be commercially marketed. As currently designed, the EMPAX instrument will meet most but not all of the FDA requirements for X-ray sources. Thus, some refinements plus any of the aforementioned enhancements must be made before this instrument could become a marketable product.

There is more emphasis today on the use of XRFS for environment monitoring and has become a recommended procedure by EPA. It is common now to use laboratory XRFS in vans for on-site analysis but the EMPAX instrument has the additional capability of performing in-situ analysis, a major advantage over sample removal and then perform the analysis.

Element	Line	Energy (keV)	Water Samples MDL, ppb*	Soil Samples MDL. ppm
Cr	Ka	5.414	100	1000
Ni	Ka	7.477	20	300
Cu	Ka	8.047	20	250
Zn	Ka	8.638	10	200
As	Ka	10.543	10	150
Se	Ka	11.221	10	140
Ag	Ka	22.162	200	>1000
Cd	La	3.133	ND	ND
Sb	La	3.605	ND	ND
Ba	La	4.467	200	>1000
Hg	Lb	11.823	20	80
Tl	Lb	12.210	30	75
Pb	Lb	12.610	20	70

ND = Not Detectable
*Assumes 100 ml sample volume containing single element only
ppm = parts per million ppb = parts per billion

Table 1. Mininum Detection Limits (MDL) for Priority Elements

Module	Size (in)	Weight (lb)	Power
Water Sample	18x18x10	40	AC
Portable Analyzer	22.5x11x12.5	34	4-Gel Cell Batt.
Electronics	4.5x9x11	12	5-NiCd Batt.
Computer	15x11x2.5	12	Rechargeable Batt.

Detector - LN2 cooled Si(Li)
X-ray Source - 30 Kv - Mo target - 0.2 ma
PHA - 4096 channels
0-600 volts bias power

Water Sample Module - Martin Marietta Design
Portable Analyzer Module - PGT Detector
- X-Tech X-ray Tube
- Martin Marietta Design
Electronics Module - Canberra Series 10 PHA
Computer Module - Grid Case 2 Portable Computer

Precipitant - Dibenzyldithiocarbamate
Filter Membrane - Rainin Nylon - 0.45 um pore size

Patent number - 5014287
NASA LaRC Contacts - Warren C. Kelliher (804) 864-4172
- Technology Utilization Office (804) 864-2482
EPA Las Vegas Contact - Bill Engelmann (702) 798-2664

Table 2. EMPAX instrument information

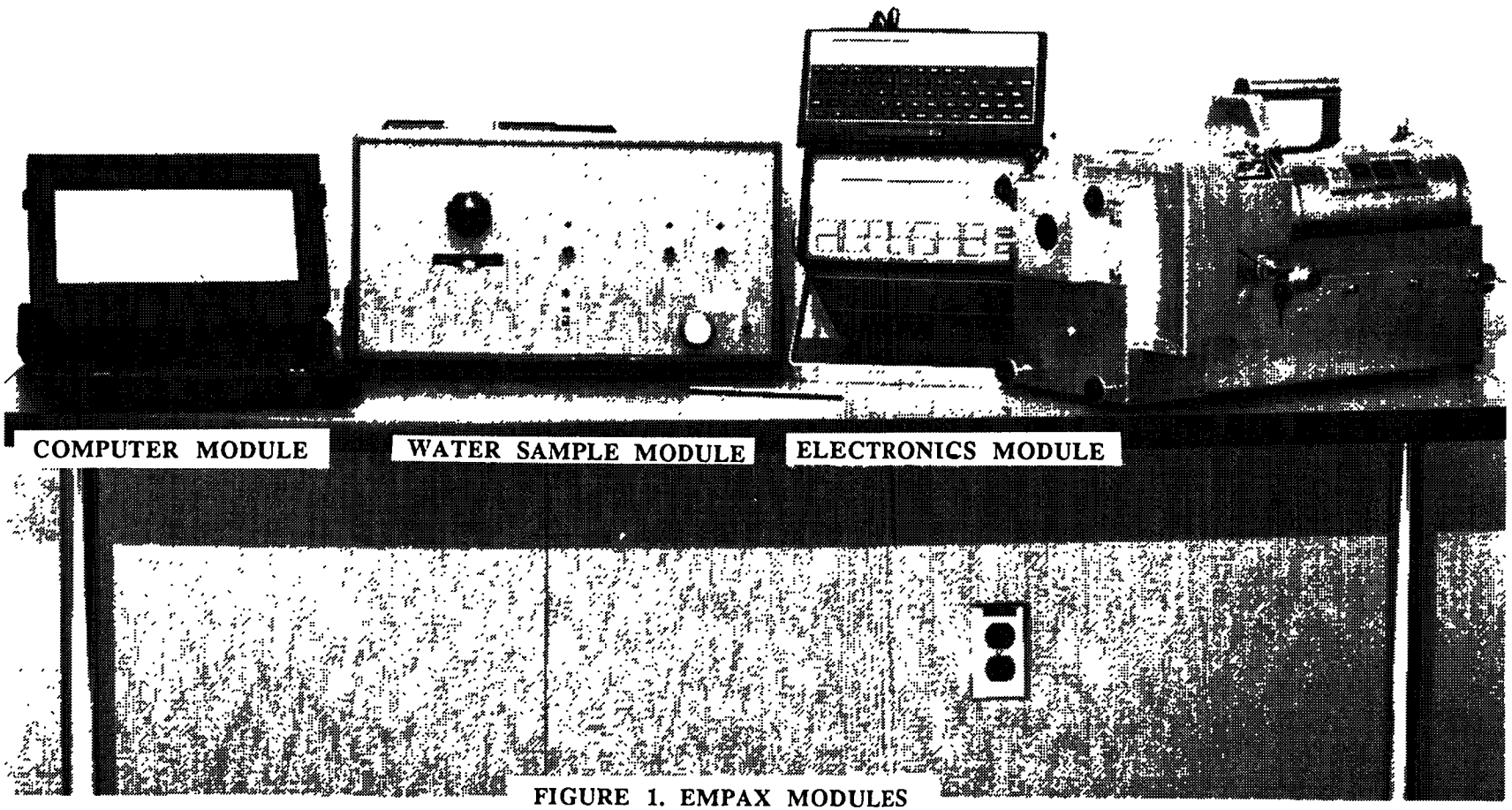


FIGURE 1. EMPAX MODULES

FIGURE 2. EMPAX MODULES WITH ANALYZER IN WATER SAMPLE MODULE

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

FIGURE 3. EMPAX INSTRUMENT BEING USED AT CONTAMINATED SITE

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

N92-22725

REMOTE SEMI-CONTINUOUS FLOW RATE LOGGING SEEPAGE METER

William G. Reay
Department of Biology
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Harry G. Walthall
NASA Langley Research Center
Hampton, Virginia 23665-5225

ABSTRACT

The movement of groundwater and its associated solutes from upland regions has been implicated in the degradation of receiving surface water bodies. Current efforts to directly measure this influx of water incorporate manually operated seepage meters which are hindered by severe limitations. A prototype seepage meter has been developed by NASA Langley Research Center and Virginia Polytechnic Institute and State University that will allow for the semi-continuous collection and data logging of seepage flux across the sediment-water interface. The meter is designed to operate at depths to 40 meters, and alleviate or minimize all disadvantages associated with traditional methods while remaining cost effective. The unit was designed to operate independently for time periods on the order of weeks with adjustable sample sequences depending upon hydrologic conditions. When used in conjunction with commercially available pressure transducers, this seepage meter allows for correlations to be made between groundwater discharge and tidal/sea state conditions in coastal areas. Field data from the Chesapeake Bay and Florida Bay systems are presented.

INTRODUCTION

Considerable research has been devoted to defining material input into inland and coastal waters. Advective transport of water and associated solutes across the sediment-water interface has been shown to be of significant importance in lacustrine, estuarine and marine environments [1,2,3,4,5]. Mechanisms responsible for advective transport of solutes within nearshore sediments include: 1) elevated upland hydraulic head (i.e., groundwater discharge) [5], 2) convective flows caused by thermal and salinity density differences [6,7,8], 3) sedimentation [9], 4) spatial variations in sea state (i.e., subtidal pumping) [10], 5) benthic boundary currents [11] and 6) bioadvection. In addition to influencing water quality management efforts, such transport mechanisms are of biological and geological importance.

Macroscopic seepage rates of water across the sediment-water interface typically range from 0.0 to 5.0 $Lm^{-2}hr^{-1}$. Hydraulic head differences between overlying surface waters and interstitial water is on the order of millimeters or centimeters for most nearshore environments [12]. Given such low and varied flow rates and hydraulic head differences, current flow meter technology is inadequate and the measurement of such phenomena remains a technical problem.

Conventional methods to directly measure water exchange across the sediment-water interface are limited to manually operated seepage meters [3]. In its simplest form, seepage meters consist of a cylinder covered by a vented lid that allows a water collection bladder to be attached. In principle, seepage meters are placed into the sediment and water entering the meter displaces water into the collection bladder. Discharge is determined by the volume of water displaced per cross sectional area of seepage meter per unit time. Discharge rates are corrected for anomalous short-term influx of water to the collection bladders [13]. Primary disadvantages of current meters include: 1) limited time series data, 2) extensive work effort required, 3) limited to safe diving depths, and 4) anomalous short-term water influx to collection bladders. The remotely operated seepage meter presented here, Sea Seep I, was designed to alleviate or minimize all disadvantages associated with traditional methods while remaining cost effective.

DESIGN AND CONSTRUCTION

The prototype seepage meter is a remotely operating self contained system consisting of seven major components, these are: 1) a magnetically controlled proximity switch, 2) motor driven system operating valve, 3) seepage collection bladder, 4) discharge adjustable metering pump, 5) data logger, 5) rechargeable battery powered electrical system, and 7) dual chamber seepage meter housing (Figures 1 and 2). Total weight of the seepage meter and components is approximately 50 kilograms.

System Operation

The main body of the prototype seepage meter is a cylinder made of 16 gauge stainless steel with dual compartments having a cross sectional area of 0.25 m^2. Batteries, system operating electrical circuit, data logger, metering pump, motorized valve and vented bladder isolation chamber are housed in the water proof upper compartment of the meter base. Compartment reenforcement allows the device to be used at depths to 40 meters. A ceramic magnet attached to the collection bladder decreases in proximity to the reed switch as the bladder fills with displaced water from the main body. When the reed switch is closed by the magnet, an electrical circuit initiates to sequentially rotate the 3-way valve from the sampling position to the input of the metering pump. The pump discharges water from the collection bladder until the proximity of the magnet increases and opens the reed switch. Subsequently, the pump is deactivated, data is logged, the motor resets the valve to the sampling position, and the system shuts down to conserve electrical power. The valve operation, pump out and data logging cycle interrupts sampling for only 15 seconds. Depending upon hydrologic conditions and data requirements, pump out volumes can be modified over a wide range by varying bladder size, proximity switch adjustment and pump cycle timer. The seepage meter is initiated and deactivated by an external magnet and internal magnetic switch.

Figure 1. Photograph of remote semi-continuous flow rate logging seepage meter.

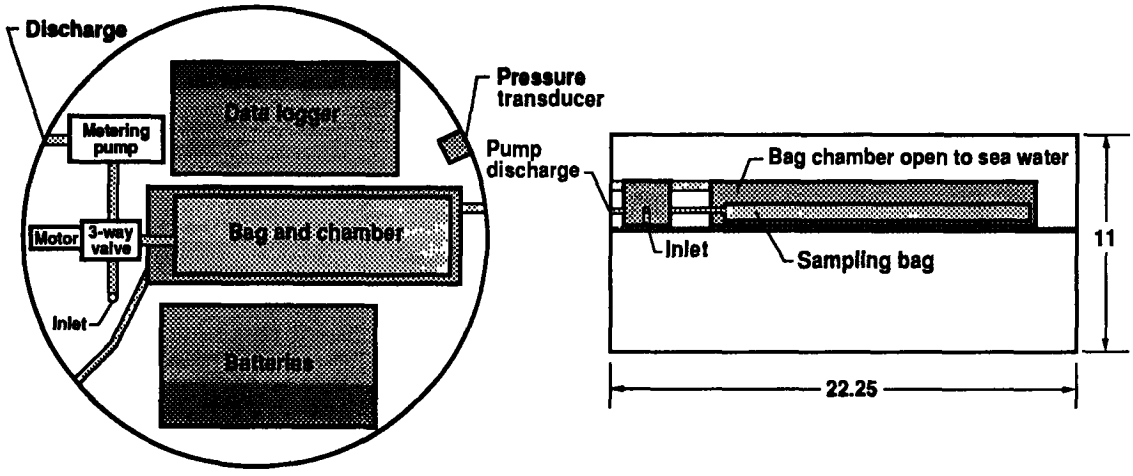


Figure 2. Schematic drawing and major components and overall dimensions of remote semi-continuous flow rate logging seepage meter.

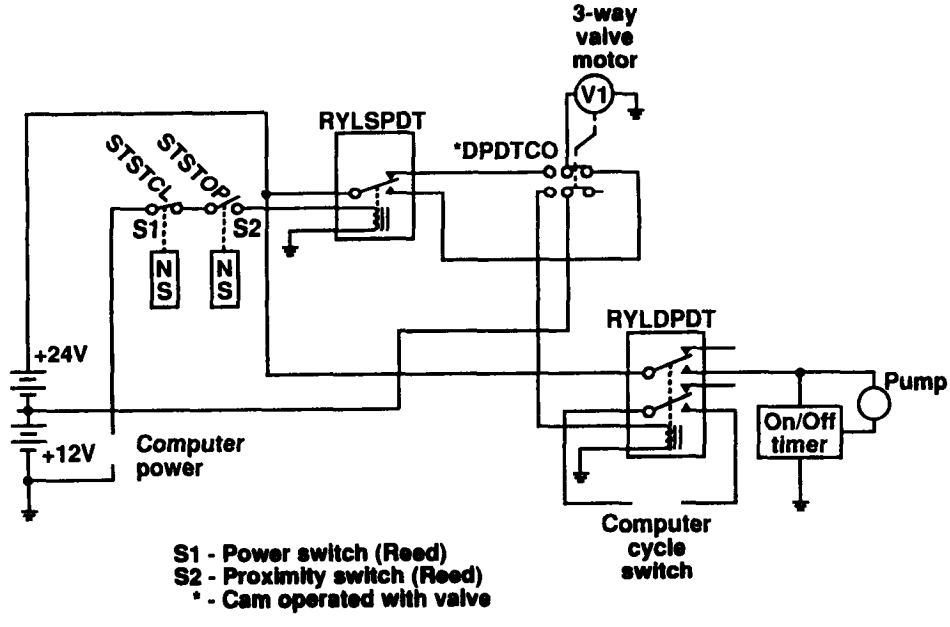


Figure 3. Electrical schematic of major circuits of remote semi-continuous flow rate logging seepage meter.

RESULTS

Calibration Tests

Due to the flow resistance caused by the components inherent to seepage meter designs and alteration of water flow paths induced by meter placement, experiments were conducted to compare actual versus measured discharge rates. Efficiency experiments were conducted in a constant head flow tank (1.0 m^2 cross sectional area) filled with a well-sorted fine sand. The constant head flow tank was allowed to equilibrate to a specific discharge for 1 hour and calibration experiments were conducted for a 2 hour period. The seepage meter was reinstalled for each individual test. Measured discharges (Q') were determined by equation (1):

$$Q' = (C_c * V)/(A * T) \qquad (1)$$

Where:

- Q' = Measured discharge ($ML^{-2}T^{-1}$)
- C_c = Cumulative counts
- V = Pump out sequence volume (L^3)
- A = Area enclosed by seepage meter (L^2)
- T = Time period (T)

The ratio of Q':Q , where Q is the actual discharge rate, was compared to determine seepage meter efficiency and accuracy [14]. Experimental results comparing seepage meter efficiency under varying discharge rates are given in Table 1. Mean efficiency of standard manual meter design has been shown to be approximately 60 percent [14]. Following initial installation, a specific time period is required to equilibrate hydraulic pressures between ambient surface water and water enclosed within the meter system, however this is generally on the order of minutes.

Table 1. Seepage meter efficiency under varying discharge rates.

Actual Discharge Rate ($Lm^{-2}hr^{-1}$)	Measured Discharge Rate ($Lm^{-2}hr^{-1}$)	Percent Mean Efficiency
6.00	1.26	21
10.32	2.52	26

Even though the seepage meter displays a slight negative buoyancy, settling of the meter into the sediment may occur resulting in an increase of pressure inside the meter and displacement of water into the collection bladder (i.e., 1 mm of settling would result in a 250 ml displacement of water). The effect of settling was determined by installing the meter into the sediment under no-flow conditions. Results for the well sorted sand indicated no significant settling effect. However, caution should be exercised in high porosity, low dry bulk density unconsolidated sediments (i.e, silt-clay mixes).

Field Application

Sea Seep I was field tested at two locations varying in sediment type and seepage discharge. The first site was located in a tidal creek on Virginia's Eastern Shore and characterized by a tidal range of approximately 1.0 meter. Nearshore surficial (upper 20 cm) quartz sandy sediments were conducive for water transport, exhibiting a mean porosity and vertical hydraulic conductivity of 0.45 and $10^{-2.4}$ $cmsec^{-1}$, respectively. The second site was located in the nearshore zone of Florida Bay which exhibits a tidal range on the order of 0.10 meters. Vertical hydraulic conductivity and porosity of surficial carbonate sediments were on the order of $10^{-2.0}$ $cmsec^{-1}$ and 0.40 respectively. Field test results for the Virginia Eastern Shore and Florida Bay sites are presented in Figures 4 and 5, respectively.

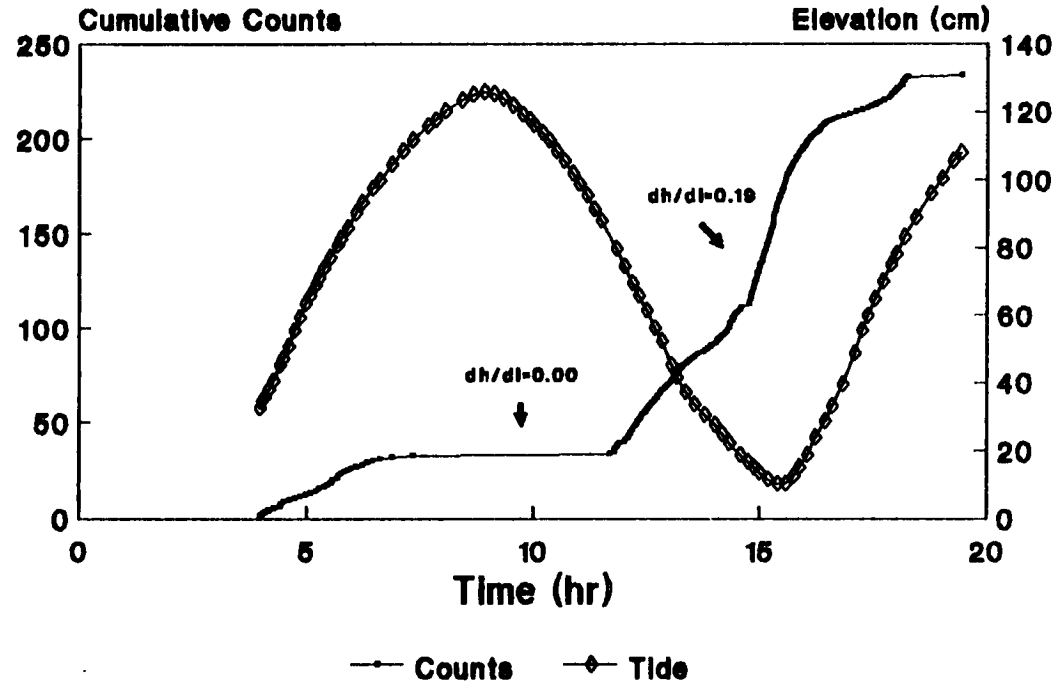


Figure 4. Graphic representation of Sea Seep I field test of tidal creek site in Chesapeake Bay.

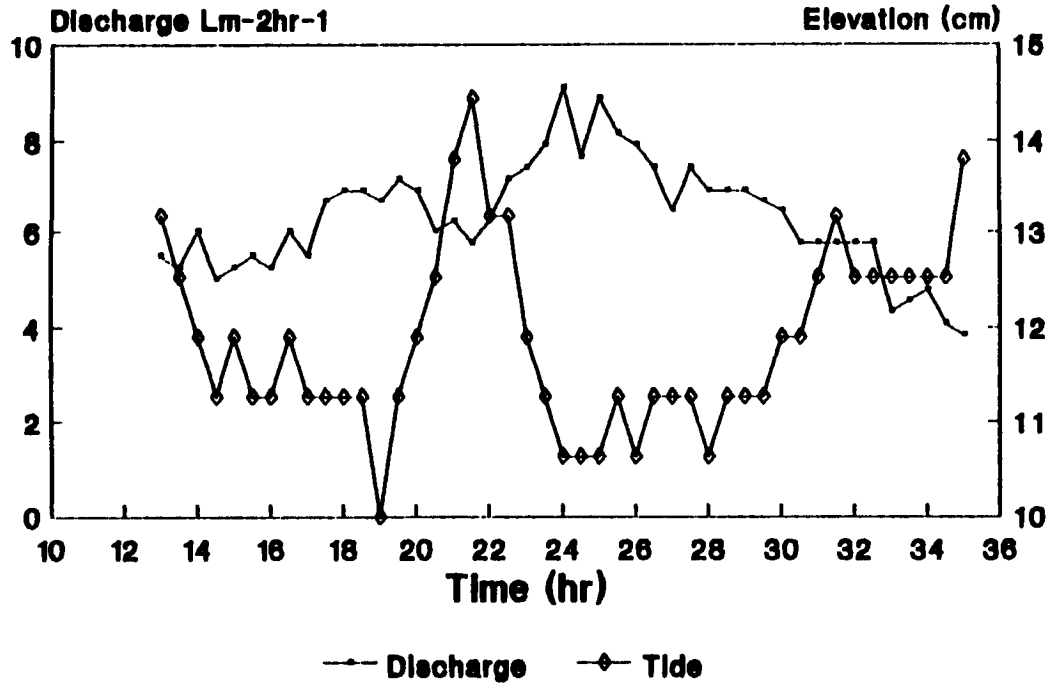


Figure 5. Graphic representation of Sea Seep I field test of nearshore site in Florida Bay.

Discharge (i.e., counts) rates show an inverse relation to tidal elevation at the Chesapeake Bay site. During high tides, water levels within the tidal creek caused hydraulic gradients between a point 0.7 meters below the sediment-water interface and surface waters to approach and equal 0.00, thereby effectively shutting down submarine groundwater discharge. Hydraulic gradient (dh/dl) is the change in hydraulic head per unit distance (l), where hydraulic head (h) describes the total energy in a moving mass of water at a particular point. Conversely, as the tide recedes, vertical hydraulic gradients (up to 0.19 at ebb tide) and discharge rates increase in concert. The remotely controlled seepage meter allowed for the collection of 246 discharge data points as compared to the 4-10 time integrated data points that would have normally been collected using manual methods. As with the Chesapeake Bay site, the Florida Bay study site displayed a strong correlation with tidal stage (Fig. 5). It should be noted that the data set for Florida Bay was summarized and consists of 1166 data points collected over a 22 hour time period.

SUMMARY AND DISCUSSION

The reported remotely operated semi-continuous flow rate data logging seepage meter was developed to, 1) operate over extended periods of time without the need for recurrent monitoring, 2) measure relatively low flow rates over short periods of time, 3) log flow rate data in addition to tidal and sea conditions, and 4) to remain cost effective. To accomplish this, the design employed a modification of the collapsible bag method used in manual seepage meters and commercially available components such as rechargeable batteries, electrical valves and pumping systems, scientific data loggers, and pressure transducers.

Development of this prototype remote seepage meter was driven by an environmental concern for inland and coastal water bodies. The data presented for the Chesapeake Bay and Florida Bay region exemplifies the geographic universality and potential importance of the phenomena, submarine groundwater discharge. Time series data allows for comparison with hydrologic conditions such as tidal influence on discharge rates.

Current uses for this instrument are generally limited to research activities and water quality management applications. This basic concept in flow meter technology does provide an alternative to applications that address extremely low flow rates in conjunction with low pressures. When used in conjunction with proper hardware material, potential industrial uses of this flow meter concept may include flow rate measurements of caustic and corrosive fluids. A second remote seepage meter has been designed with modifications that increase meter efficiency and sensitivity, and allow for deeper water deployment.

We are proceeding to protect this technology with a patent.

ACKNOWLEDGEMENTS

Numerous friends and colleagues made significant contributions throughout the development, calibration and field testing phases of Sea Seep I. This project could not have been possible without the guidance and high morale of Dr. George Simmons, Jr. of VPI&SU. The authors would like to thank Dr. Daniel Gallagher of VPI&SU and Gary Clow of the U.S. Geological Survey for providing their expertise with regards to the data storage and access aspects of this project. Fabrication was performed at NASA Langley Research Center with the technicial expertise and assistance of Gene Fullerton, Noland Liddle and Earl Knight of the Metals Technology Development Section, Mark Simonton and Ellis Hogge of the Advanced Machining Development Section, Dennis Noe and James Bell of the Electronics Fabrication Development Section. We also thank John Samos, Joseph Mathis and Dr. Franklin Farmer at NASA Langley's Technology Utilization and Applications Office for continued support throughout the development phase of this project. We also gratefully acknowledge Dr. Alan Hulbert and Dr. Steven Miller of NOAA's Undersea Research Center/UNC-Wilmington, for their support during initial field testing. We are grateful to Jake Waller, Scott Smedley and Eduardo Miles for support during field testing, and Karen Reay and Mark Kidd of VPI&SU for assistance during calibration experiments. A special thanks is given to Hartmut Beck of the Fachhochschule Hamburg, who spent numerous hours investigating various concepts and low flow rate sensor designs while on internship at VPI&SU. We would also like to acknowledge Jim Sherrill, who provided technical support during the early stages of our endeavor.

REFERENCES

1. Belanger, T.V. and D.F. Mikutel. 1985. On the Use of Seepage Meters to Estimate Groundwater Nutrient Loading to Lakes. *Water Resources Bulletin* 21(2): 265-272.

2. Johannes, R.E. and C.J. Hearn. 1985. The Effect of Submarine Groundwater Discharge on Nutrient and Salinity Regimes in a Coastal Lagoon off Perth, Western Australia. *Estuarine, Coastal and Shelf Science* 12: 789-800.

3. Lee, D.R. 1977. A Device for Measuring Seepage Flux in Lakes and Estuaries. *Limnology and Oceanography.* 22: 140-147.

4. Lewis, J.B. 1987. Measurements of Groundwater Seepage Flux onto a Coral Reef: Spatial and Temporal Variations. *Limnology and Oceanography* 32(5): 1165-1169.

5. Simmons, G.M., Jr. 1989. The Chesapeake Bay's Hidden Tributary: Submarine Groundwater Discharge. pp. 9-28. In: *Proc. Ground water Issues and Solutions in the Potomac River Basin/Chesapeake Bay Region*. Washington, DC: Co-sponsored by the Assoc. of Ground Water Scientists, et al., and George Washington University.

6. Kohout, F.A. 1965. A Hypothesis Concerning Cyclic Flow of Salt Water Related to Geothermal Heating in the Floridian Aquifer. *New York Academy of Science Series II*, 28: 249-271.

7. Sayles, F.L., and W.M. Jenkins. 1982. Advection of Pore Fluids Through Sediments in the Equatorial East Pacific. *Science* 217: 245-248.

8. Simms, M. 1984. Dolomitization by Groundwater-Flow Systems in Carbonate Platforms. *Transactions of the Gulf Coast Association Geologic Society.* 34: 411-420.

9. Tzur, Y. 1971. Interstitial Diffusion and Advection of Solute in Accumulating Sediments. *Journal of Geophysical Research* 76(18): 4208-4211.

10. Riedl, R.J., N. Huang, and R. Machan. 1972. The Subtidal Pump: A Mechanism of Interstitial Water Exchange by Wave Action. *Marine Biology* 13: 210-221.

11. Thibodeaux, L.J. and J.D. Boyle. 1987. Bedform Generated Convective Transport in Bottom Sediment. *Nature* 325: 341-343.

12. Reay, W.G. and G.M. Simmons, Jr. Personal Communications. Dept. of Biology, VPI&SU, Blacksburg, Va.

13. Shaw, R.D. and E.E. Prepas. 1989. Anomalous, Short-term Influx of Water into Seepage Meters. *Limnology and Oceanography* 34(7): 1343-1351.

14. Cherkauer, D.A. and J.M. McBride. 1988. A Remotely Operated Seepage Meter for Use in Large Lakes and Rivers. *Ground Water* 26(2): 165-171.

N92-22726

CALCIFICATION PREVENTION TABLETS

Geoffrey A. Lindsay and Michael A. Hasting
Chemistry Division, Research Department
Naval Weapons Center, China Lake, CA 93555
and
Michael A. Gustavson
NAVSEADET (PERA CV)
Bremerton, WA 98310-4924

ABSTRACT

Citric acid tablets, which slowly release citric acid when flushed with water, are under development by the Navy for calcification prevention. The citric acid dissolves calcium carbonate deposits and chelates the calcium. For use in urinals, a dispenser is not required because the tablets are non-toxic and safe to handle. The tablets are placed in the bottom of the urinal, and are consumed in several hundred flushes (the release rate can be tailored by adjusting the formulation). All of the ingredients are environmentally biodegradable. Mass production of the tablets on commercial tableting machines has been demonstrated. The tablets are inexpensive (about 75 cents apiece). Incidences of clogged pipes and urinals were greatly decreased in long term shipboard tests. The corrosion rate of sewage collection pipe (90/10 Cu/Ni) in citric acid solution in the laboratory is several mils per year at conditions typically found in traps under the urinals. The only shipboard corrosion seen to date is of the yellow brass urinal tail pieces. While this is acceptable, the search for a nontoxic corrosion inhibitor is underway. The shelf life of the tablets is at least one year if stored at 50% relative humidity, and longer if stored in sealed plastic buckets.

INTRODUCTION

This project grew from the problem of calcium carbonate buildup in the collection, holding and transfer (CHT) piping (i.e., the ship's sewer system) and plugged urinals on the U.S. Navy's aircraft carriers. Sea water is used to flush toilets on ships. When sea water mixes with urine, it precipitates calcium salts which are insoluble and form hard deposits when the pH of the water is above about 6 (sea water has a pH of about 8.5 depending on the dissolved gases). Acids have been shown to dissolve the calcium deposits in the sewage piping. The treatment used on aircraft carriers has been to place in the urinals a perforated plastic bag of sulfamic acid powder held in a plastic dispenser. It is only marginally effective in removing calcification in the pipes. For aircraft carriers, nearly one million dollars annually is consumed with hydroblasting of clogged pipes. There are other costs associated with fleet down-time which are not easily quantified. Sulfamic acid powder is a very strong acid; hence, it presents an irritation and potential safety hazard to the user. It is difficult to determine when the bags of sulfamic acid are depleted, and the sailors dislike removing the used bags from the plastic dispensers in the urinals.

This paper describes the three-year development and testing of citric acid tablets, tablet manufacturing trials and shipboard decalcification trials. The shipboard tests had an immediate impact on the sailors by reducing the incidence of plugged urinals.

TABLET DEVELOPMENT

Design Criteria. We set the following requirements for developing calcification prevention tablets. The ingredients should be: (1) water soluble, (2) biodegradable, (3) nontoxic, and 4) commercially available. The tablets should: (5) be manufacturable on commercially available equipment by more than one company, (6) give a slow, controlled rate of acid release, (7) be mechanically strong, (8) be low cost, and (9) not require a dispenser.

Attempts to bind sulfamic acid into a tablet were abandoned in favor of citric acid for a number of reasons: (1) citric acid is more environmentally acceptable, (2) citric acid is a chelating agent for calcium, (3) citric acid tablets have superior mechanical integrity, and 4) the total cost is essentially equivalent.

Comparison of Water-Soluble Binders. A large number of synthetic and natural water-soluble polymer binders, including poly(vinyl alcohol), carboxymethyl cellulose, xanthan gums, were investigated in the laboratory and found to be unsuitable due to poor binding or poor release qualities. We found that polyethylene oxide (PEO) gave the best balance of dissolution rate control, manufacturability, availability, and environmental acceptance.

Other Additives. The kinds of additives which we believe are important are: 1) high molecular weight (PEO) binders (for release rate control), 2) compatible processing aids (calcium stearate), 3) hardness control additives (sorbitol and PEG), 4) desiccants (fumed silica), and 5) corrosion inhibitors. We are currently screening corrosion inhibitors for 90/10 Cu/Ni pipe. We recommend excluding coloring agents, perfumes and deodorants. Coloring could splash and stain white Navy uniforms. Odors, which could indicate deteriorating vent piping which must be repaired to prevent dangerous accumulations of hydrogen sulfide gas, should not be masked.

Selecting the Formulation. The present preferred tablet formulation, which provides a pH of about 4.0 to 4.5 in the shipboard urinals, is: citric acid = 70%; PEG = 16%; Sorbitol = 10%; PEO = 3.5%; Ca stearate ≤ 0.25%, and fumed silica ≤ 0.25%

Biodegradability of Tablet Ingredients. Citric acid, calcium stearate,[1] and sorbitol[2] are readily biodegradable (>60% in 10 days). The solubility of calcium stearate is 2 mg/l which is sufficient to ensure availability to bacteria. Polyethylene glycol (PEG) biodegrades slowly.[3] Polyethylene oxide (Polyox^R), a higher molecular weight form of PEG, will also biodegrade with time. Silica in many forms, including quartz, can be metabolized by biological action.[4] For example, several bacteria and plants can produce monomer silica from quartz and other solid (polymer) forms of silica. The resulting monomer silica is taken up by various life forms (eg., diatoms) and used for skeletal material. The lethal oral dose of sulfamic acid for rats is 1.6 g/kg.[5]

Solubility. The saturation solubility of citric acid in water at 10 °C (50 °F) is about 54 grams per 100 grams of solution, and at 80 °C (176 °F), about 79 grams per 100 grams of solution. Hence, dissolved citric acid is available in high concentrations. The chelation property of citric acid causes it to bind tightly to calcium ions. The solubility of calcium citrate in sea water is low; however, the insoluble calcium citrate is a soft, hygroscopic solid which is easily suspended and flushed out of the system.

pH Control. The pH determines the kinetics of calcium dissolution and pipe corrosion more than any other variables, such as temperature and buffering ingredients. We have found that citric acid has a dampening effect on the pH swings. Whereas, in dissolution tests with sulfamic acid, the pH jumps up and down to a much greater extent. The same spiked pH phenomenon was observed in shipboard field tests with sulfamic acid.

TABLET PERFORMANCE TESTING IN THE LABORATORY

Controlled Release Rate of Acids. A laboratory test for measuring the dosage of acid released from the tablets under controlled conditions which simulate actual shipboard urinal conditions was developed. The laboratory dissolution rate tester, shown in figure 1 on the next page, evolved from several earlier prototypes. The release rate of citric acid was determined by recording the pH in the 0.7-liter overflow vessel until the tablet was essentially dissolved. Tap water or sea water[6] (0.5 liters/flush) was used. The flow rate was adjusted to give about one flush every 60 seconds. The temperature of the water was between 75 and 85° F. The tablets should maintain sea water between a pH of about 3.2 ± 0.2 to about 5.4 ± 0.2 (in the 0.7-liter overflow container) for about 250 ± 50 flushes to meet our specifications.

Citric acid tablets were made with various amounts (and various molecular weights) of PEO, PEG and citric acid to find compositions which gave desirable release rates and mechanical integrity. At constant PEO to citric acid ratio, the release rate was essentially unchanged by sorbitol, the hardening agent. After considerable testing, we felt that 3.5% PEO would be the most effective composition in controlling the calcification of the shipboard CHT lines. From all the data collected, both in the lab and on the ship, a tablet formulation was selected for scale-up. The main purpose of the scale-up runs were to determine if citric acid tablets could be mass produced. Several manufacturers have now successfully produced several hundred thousand tablets which are three inches in diameter and about one-inch thick (145 grams of ingredients). This tablet size is about as large as can be conveniently manufactured on commercial equipment, and the tablet gives two to three days of service in heavily used urinals.

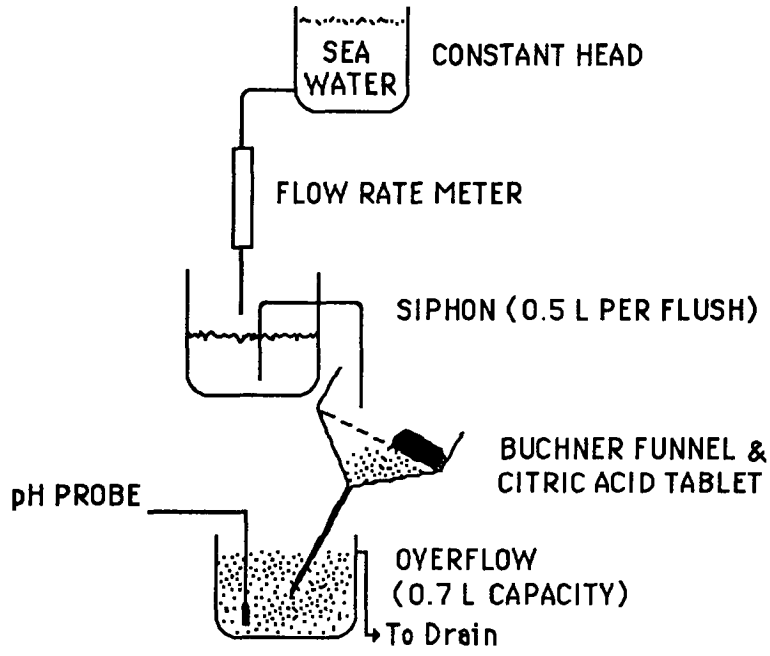


Figure 1: Laboratory Dissolution rate tester

TABLETING PROCESS DEVELOPMENT

Comparison of Various Processes. The process of making the tablets went through several stages. At first citric acid was mixed with the binders at room temperature, heated in a mold above the melting point of the binders, placed in a hydraulic press and cooled under pressure. This worked well, but we found that making tablets by room temperature compression was a more efficient process. It was discovered that combining a mixture of the ingredients as powders into a die and subjecting it to several tons of pressure, the tablets would retain shape in water and slowly dissolve at desirable rates.

Effect of Temperature and Humidity in the Manufacturing Process. It was found that high humidity greatly reduces the ability to manufacture high quality tablets. Relative humidity in the tableting room should be keep below 25%. If the relative humidity is too high, citric acid, which is anhydrous, will attract moisture from the atmosphere. In the press, the moisture is squeezed out of the citric acid which makes the binder tacky. Although dry-to-the-touch, under the pressure of the tableting process, the tablets may stick to the die.

The addition of a small amount of a desiccant to the tablet aids the binder tableting process, eg., about 1/4% or less of fumed silica. In addition to this, about 1/4% or less of calcium stearate is added to act as a die-lubricant and to help with the sticking problem. After the tablet is sealed in its plastic wrapper it takes several months of exposure to humidities above 75% before the tablet is effected by the moisture. After prolonged exposure to high humidity, the tablet softens, but its performance in the urinal is unaffected. The plastic-wrapped tablets are stored in a polyethylene bucket which is sealed from the environment. Even though the tablets are manufactured and packaged in a very low humidity environment, we believe a desiccant pack should be added to the container in order to absorb any moisture that may get into the container if the lid is not properly resealed.

CITRIC ACID TABLET PERFORMANCE

Reactions with calcium carbonate. After two months of treatment with citric acid tablets in shipboard tests, it was found that hard build up in the pipes was greatly softened, and the soft build up was very easy to remove by flushing. Although it has not yet been analyzed, the soft build up is likely to be calcium citrate based on the following. Citric acid reacts with calcium carbonate giving off carbon dioxide and various forms of calcium citrate. The solubility of tricalcium citrate tetrahydrate ($Ca_3(C_6H_5O_7)_2$ plus 4 moles of water of crystallization) is reported

to be 0.85 grams per 100 cc water at 18° C and less at higher temperatures.[7,8] In the presence of excess citric acid, the tricalcium salt is in equilibrium with the dicalcium hydrogen salt which is thermodynamically preferred[9] in warmer water: $Ca_3(C_6H_5O_7)_2$ + citric acid <=> $CaH(C_6H_5O_7)$.[10] The dicalcium hydrogen salt may be more soluble, but no data on its solubility has been found. The pH and other buffering agents will also play an important role in the kinetics of calcium carbonate dissolution and calcium citrate precipitation. To learn more about solubilities, the following tests were run.

We obtained samples of calcified solids which had formed in the CHT pipes of aircraft carriers when the ship had been using only sulfamic acid bag treatment. (The citric acid tablets had not been invented when these samples were obtained.) The solids are mostly calcium carbonate, but are also thought to contain calcium oxalate, uric acid, and other calcium salts.[11] About one gram of the calcified solid was placed in 100 ml of a seawater-solution containing various amounts of the acid to be tested. The time to completely dissolve the calcified solid was measured. The tests in our laboratory at room temperature, showed that one or more part of citric acid per part of solid dissolved the solid in a few minutes giving off bubbles (CO_2) and leaving behind a brown cloudy solution.

The results from controlled experiments using pure materials are shown in Table I. The amount of pure $CaCO_3$ was held constant at 1.38 grams (0.01235 moles) per 100 cc tap water at room temperature.

Table I. Dissolution of Calcium Carbonate at Various concentrations of Citric Acid

Test #	Acid type	Acid/Ca (mole ratio)	Comments
1	pure citric	0.74	solution did not became clear
7	pure citric	1.01	clear in 30 min.; ppt. in 60 min.
10	pure citric	1.30	clear in 7 min.; ppt. in 25 min.
9	pure citric	1.56	clear in 4 min.; ppt in 2 hours.
12	pure citric	2.08	clear in 3 min.; ppt in 2 days.

It would appear that citric acid is effective in dissolving hard calcium deposits, and forms some fine particles suspended in water at these concentrations (probably calcium citrate) which could be flushed through the pipes into the holding tank.

Mechanical Properties of Tablet. A quick and dirty test was devised which simulates conditions likely to be encountered in the field. Tablets are dropped on a hard floor (vinyl tile-covered concrete) from a three-foot elevation. If the tablet does not break, crack or chip or produce more than 2% loose powder, it passes.

A more scientific test will be initiated, if problems arise. One such test would be to support the tablet with two parallel bars and drop a heavy dart from known heights, increasing the height until the tablet breaks. The energy to break the tablet can be determined statistically (eg., the foot-pounds of force at which 50% of the tablets from one lot survive).

Long Term Aging. Several lots of citric acid tablets, in their Mylar^R wrapper, were stored at room temperature in a plastic bucket with its lid cracked open, thus exposing tablets to variable humidity conditions in a Los Angeles warehouse. After one year, they were tested for dissolution rate vs pH, and breakage resistance. The flushes per tablet for these 3% PEO formulations ranged from 180 to about 220 for the year-old tablets; whereas, the fresh tablets lasted about 240 to 280 flushes. However, it should be noted that the temperature of the flush water used for the aged tablets was 30° C compared to only 23° C for the fresh tablets. The temperature difference could account for most of the difference in number of flushes. The tablets passed the breakage test.

A dozen wrapped tablets (3% PEO), were stored in a chamber at a constant 50% relative humidity at room temperature for one year. After one year, the tablets were noticeably softer (fingernail indentation test). The dissolution rate ranged from 300 to 350 flushes over the required 3 to 5.5 pH range using 28° C tap water. These data would indicate that the tablets have a shelf-life of at least one year. We speculate that exposing the tablets to humid air over long periods may decrease the dissolution rate due to "solvent-welding" of the citric acid particles with the binder, so that fewer discrete particles slough off the tablet, giving more time for the molecular dissolution of citric acid from the surface of the tablet.

PACKAGING CONSIDERATIONS

Because the tablets are water soluble and most storage aboard ship is very humid, care must be taken in the packaging of the tablets. To keep the tablets dry, each tablet is sealed in a plastic wrapper then placed in a five gallon plastic bucket. Theses buckets must be water tight, and in each bucket a desiccant package is added. A polyethylene bucket with an O-ring seal was chosen. After the bucket is opened aboard ship, this package of desiccant will absorb some of the moisture in the air, but is only effective if the bucket is sealed again with its original lid in a timely manner.

Even though the plastic wrapping acts as a barrier to water and moist air, it is understood that a small amount of water will diffuse through any plastic film over long periods of time. However, the buckets are the first line of defense against moisture. The wrapping is also gives mechanical protection to the tablets during packing, shipping and unloading of the buckets. We have had success with polyester (Mylar^{R}), polyethylene and related copolymers.

SHIPBOARD TESTING

From February, 1989 until present, many tests have been run in USS Independence (CV 62). In the first test aboard USS Independence, a bank of four urinals was used to test the pH and dissolution rates of the tablets. The pH in the urinals was measured as a function of various formulations of the citric acid tablet. Readings ranged from as low as 3 to around a high of 6. One could see when the ships crew got up in the morning and when they went to bed at night by the change in pH in the CHT lines. In the lines down stream from the urinals, the pH would drop as a urinal was flushed. But with the addition of the toilets' flow of water (about 15 times more volume), the pH would again rise to over seven.

A six month supply of the 3%-PEO formulation was given to USS Independence in March, 1990. The ships trouble logs show the rate of call for clogged urinals went from about four per day to about one per day after this supply of tablets was used by the ship. This proved to be of great benefit to the ships maintenance personnel as those sailors normally assigned to fixing the drains could now be used in other places on the ship. With the use of the 3% PEO tablets, enough citric acid was added to soften the calcification in the main pipes.[12] Several valves were pulled after the return of USS Independence from the Persian Gulf, and little-to-no calcification was found in the CHT system. The pipes had been cleaned by hydroblasting at least two years earlier and should have had major problems with the CHT system from calcification. Tests in USS Independence showed a usage rate of about 0.027 tablets per day per sailor, this number fits well with the calculated amount of tablets before the test began.

More controlled tests have been done in USS Nimitz (CVN 68) where valves have been pulled before the use of the citric acid tablets and the calcification measured and photographed in October, 1990 and February, 1991. After only two months of citric acid tablet use, softening of the calcification and removal of some of the calcification had taken place.

Response from the aircraft carriers has been very favorable, with the only problems being corrosion of the tail pieces of the urinals. The yellow brass tail pieces can be periodically replaced or a fiberglass tail piece can be used. We are also investigating adding corrosion inhibitors to the tablet formulation.

CORROSION TESTING OF 90/10 COPPER/NICKEL PIPE

The Naval Civil Engineering Laboratory (NCEL) was tasked to compare corrosion rates of 90/10 copper-nickel metal exposed to: (1) sea water as a control, (2) sea water containing a citric acid tablet, and (3) sea water containing sulfamic acid bag contents (NSN 9G-6850-01-150-4921). This test consisted of a placing 90/10 Cu/Ni coupons in sea water at 35°C (95°F) containing a given amount of acid and monitoring weight loss of the metal coupon with time. For the most severe case (a concentration of 2.9 g/L of each acid formulation), after two weeks, the corrosion rate for the citric acid tablet (70% citric acid) was 60 mpy, and that for sulfamic acid bag (95% sulfamic acid) was 80 mpy. The pH during these tests increased with time from about 2 to about 5 for both acids. The report concludes that, at the most severe condition, corrosion due to citric acid tablets was about the same as that for sulfamic acid bag contents (NSN 9G-6850-01-150-4921). The results of this test were reported in reference.[13]

We ran another corrosion test in which the pH was held relatively constant for three weeks by adding more acid as it was consumed (as indicated by a rise in pH). The test protocols used are described in ASTM G 1-90 and ASTM G-31-72. A summary of the data are shown in the table below. The smaller corrosion rates in the NWC test compared to the NCEL test probably results from the lower temperature of the water (23° vs. 35° C).

Table II. Corrosion Rate of 90/10 Cu/Ni pipe in Various Acidic Solutions

Fluid Id.	Solution Type	pH	Corrosion Rate: mil/yr.	Comments (final color of water)
R-010042	tap water	8	0.01	washed + scrubbed 7X
Q-283240	synth. sea water	8	0.47	light blue, ppt
C-143637	pure sulfamic acid	3	7.25	light green
E-434651	sulfamic acid baggie	3	6.20	green / yellow
A-161722	pure citric acid	3	29.5	dark green
I-213547	citric tablet	3	21.9	dark green
O-263139	citric tab + Na silicate	3	20.5	dark green
M-124550	citric tab + Na molybdate	3	20.3	dark green
K-495354	citric tab + benzotriazole	3	0.72	tan / green
D-070910	pure sulfamic acid	5	2.04	light green
F-030008	sulfamic acid baggie	5	1.85	yellow
B-131819	pure citric acid	5	1.54	green
J-253338	citric tablet	5	1.63	blue / green
P-042029	citric tab + Na silicate	5	1.33	blue / green
N-020523	citric tab + Na molybdate	5	2.36	blue / green
L-152430	citric tab + benzotriazole	5	0.03	tint of green

Three 1"x3" strips of 90/10 Cu/Ni pipe cut from a 5"-dia. schedule 40 pipe were submerged without stirring in Bio-Sea Marinemix[5] (simulated sea water) at room temperature. The metal coupons were removed from the solution six or seven times for cleaning and weighing during the three-week test period. The corrosion rate at a pH of 3 is considerably higher than at a pH of 5 for all solutions. In order to hold the pH constant during the test, the solutions had to be mixed with concentrates of the acids periodically. This procedure resulted in the pH cycling to a somewhat higher value as the acid was consumed, then dropping to the desired value when the concentrate was added.

At the most severe conditions (pH = 3) citric acid corrodes at a higher rate than sulfamic acid, but at the milder conditions (pH = 5), the two acids are comparable. Measurements of pH on aircraft carriers in the sewage collection lines reveal pH ranges from 4 to 5 in the traps under the urinals and up to about 7 in the large mains. Yet the hard deposits in the mains had been reduced to soft, flushable deposits. Therefore, the citric acid treatment has been recommended for continued use.

Several corrosion inhibitors for citric acid tablets were tested. Note that benzotriazole (BTA) is an excellent corrosion inhibitor for citric acid even at the pH of 3 condition. Unfortunately, BTA is toxic: oral LD_{50} for rats is 560 mg/Kg; dermal LD_{50} for rabbits is 2g/Kg; and for sunfish, trout and minnows, the lethal dose is about 35 mg/l.[14] **Hence, we do not recommend putting BTA in the tablets.** A National Cancer Institute report on feeding benzotriazole (BTA) to rats concluded that it was not carcinogenic in Fischer 344 rats, and there was inconclusive evidence in B6C3F1 mice.[15] The sodium silicate and sodium moybdate appear to have no effect on corrosion rate by themselves.

ACKNOWLEDGEMENT

This Project was sponsored by COMNAVAIRLANT and COMNAVAIRPAC, and is managed by NAVSEADET (PERA CV). The skillful assistance of Ms. Deborah Paull, Mr. Michael Pietrak and Dr. Robert Kubin is gratefully acknowledged for the NWC corrosion study and several of the dissolution studies.

1 De Morsier, et al., *Chemosphere*, 16(4), 833 (1987).

2 Perlman, *J. Chem. Ed.*, 36, 60 (1959).

[3] Mark Reid, Union Carbide communication, Sept. 30, 1991.

[4] A.M. Lauwers and W. Heinen, *Arch. Microbiol.*, 95, 67 (1974).

[5] *The Merck Index*, P.G. Strecher, Ed., Merck & Co., Inc., Rahway, NJ, 1960; p.995.

[6] For synthetic sea water, we used Bio-Sea Marinemix (available from California Aquarium Supply House, San Carlos, CA), a dry mixture of salts mixed with fresh water to a specific gravity of 1.020 to 1.023 at 73 to 77° F. It contained the following chemical abundances (in g/kg): chloride = 18.9; sodium = 10.5; sulfate = 2.47; magnesium = 1.27; calcium = 0.40; potassium = 0.38; bicarbonate = 0.14; bromide = 0.065; boric acid = 0.024 strontium = 0.008; silicic acid = 0.003; fluoride = 0.0013; and many other trace elements.

[7] *The Merck Index*, P.G. Strecher, Ed., Merck & Co., Inc., Rahway, NJ, 1960; p.192.

[8] *Handbook of Chemistry and Physics*; R.C. Weast and M.J. Astle, Eds.; CRC Press; Boca Raton, FL, 1980-1981; p. B-87.

[9] Sometimes there is not enough time to reach the thermodynamically preferred compound, and kinetics dictates which form is found.

[10] A. Abou-Zeid and M.A. Ashy, *Agricultural Wastes* 9, 51 (1984).

[11] Private communication from Dr. Zahid Amjad, BF Goodrich Corp., Brecksville, Ohio.

[12] By Direction Ltr. 9593 Ser 042B/18 of 24 Jan 90 from NAVSSES, Philadelphia to NWC, China Lake, "Citric Acid Test on USS Independence (CV-62)."

[13] D.E. Pendleton, Test Report: "Corrosion of 90/10 Copper Nickel Metal," by direction Ltr. Ser L52/ 0253 of 13 Mar 91.

[14] MSDS from Mobay Chemical Co. (manufacturer of BTA).

[15] Technical Report No. NCI-CG-TR-88; National Cancer Inst., Bethesda, MD Carcinogenesis Testing Program; DHEW/PUB/NIH-78-1338, March 1978.

N92-22727

AUTOMATED CARBON DIOXIDE CLEANING SYSTEM

David T. Hoppe
NASA Marshall Space Flight Center
MSFC, AL 35812

ABSTRACT

Solidified CO_2 pellets have been proven to be an effective blast media for the cleaning of a variety of materials. CO_2 is obtained from the waste gas streams generated from other manufacturing processes and therefore does not contribute to the green house effect, depletion of the ozone layer, or the environmental burden of hazardous waste disposal. The system is capable of removing as much as 90% of the contamination from a surface in one pass or to a high cleanliness level after multiple passes. Although the system is packaged and designed for manual hand held cleaning processes, the nozzle can easily be attached to the end effector of a robot for automated cleaning of predefined and known geometries. Specific tailoring of cleaning parameters are required to optimize the process for each individual geometry. Using optimum cleaning parameters the CO_2 systems were shown to be capable of cleaning to molecular levels below 0.7 mg/ft^2. The systems were effective for removing a variety of contaminants such as lubricating oils, cutting oils, grease, alcohol residue, biological films, and silicone. The system was effective on steel, aluminum, and carbon phenolic substrates.

INTRODUCTION

A high level of cleanliness is a typical requirement for the sensitive hardware routinely lifted into space. The fairings surrounding this equipment must be equally clean to prevent contamination of the payload. The high vacuum of space will cause even a minute amount of residual contamination to off gas and potentially redeposit on the payload compromising its intended service.

The high speed which space craft accelerate through the atmosphere can have a devastating effect on the ship's thin skin. The friction of the atmosphere causes severe erosion and elevated temperatures. The space craft is protected form this harsh environment by applying a coating of material designed to be eroded away during the flight called an ablator. This ablative material must adhere tightly to the surface of the space craft. If the ablator were to peel off during flight, the exposure to the space craft could greatly compromise the integrity of the vehicle. The level of cleanliness is also an important factor in adhesion strength and determines to a large extent the quality of the bond.

BACKGROUND

Most standard cleaning processes utilize a labor intensive manual effort and some require large amounts of environmentally hazardous and chemically dangerous cleaning agents. Typical processes include high pressure water spray, various soaps, vapor degreasing, mechanical grit blasting, aqueous solutions, and solvents such as Freon 113, methyl ethyl ketone, isopropyl alcohol, and methylene chloride. All of theses process generate more waste than the contamination being removed. Often this waste is considered hazardous and poses a potential threat to the environment. Many of the processes require personnel to wear extensive safety protective clothing. Some chemicals such as Freon 113 have a deleterious effect on the ozone layer. The Environmental Protection Agency (EPA) has imposed reduction requirements on the use of chloroflurocarbon (CFC) compounds. Based on 1986 levels, the use of CFC's must be reduced by 50% by 1991 and eliminated by 2000.

Manual hand cleaning of components require a great deal of effort and time. Cleaning the first payload fairing of the Titan IV rocket required approximately 11,000 man hours. The processes used large quantities of Freon 113, hand wipes and cotton swabs. Hand cleaning often requires extensive cleanliness inspections and recleaning.

Clearly there is a need for a cleaning method which is radically different from those previously described. This process should greatly reduce the amount of hazardous waste generated in the existing procedures. The requirement for protective clothing and personal protection equipment should be reduced. The time involved in performing the cleaning operation should also be greatly reduced. The manpower level effort should be much lower. The process should be safe for the hardware being cleaned. The total costs due to reduction in manpower level, time expenditures, waste processing, and incorporating safety measures should be significantly reduced.

One cleaning method incorporating all these requirements was investigated in a joint venture between the USAF and NASA. The process uses a solidified CO_2 pellet media blast. This investigation was initiated by Martin Marietta Corporation (MMC) as an Industrial and Modernization Improvement Plan (IMIP) to demonstrate that the CO_2 blasting process could enhance the Titan IV Payload Fairing (PLF) cleaning process, reduce costs and meet the EPA requirements.

After an initial study of two separate CO_2 cleaning systems, one system was chosen for intensified studies and installed temporarily at MSFC. The second system was brought in later and used for a comparative study for use on cleaning components associated with the current space shuttle system and future advanced solid rocket motor materials. The CO_2 cleaning equipment and operator were provided by Environmental Alternatives. The two CO_2 cleaning systems were provided by Alpheus and Cold Jet.

CO_2 CLEANING

Process Description

Equipment

The CO_2 cleaning system is shown in figure 1. A tank stores liquid CO_2 at 200-300 psig. The pelletizer has a 2 step process. First the pelletizer transforms the liquid CO_2 into a snow like solid by quickly reducing the pressure. Next the snow is compressed and extruded into pellets approximately ¼" in diameter and ½" long. By changing the size of the die, used to shape the pellets, a wide variety of pellet sizes can be achieved. The system must be shut down to change die sizes.

A compressed oil free air source at 400 psig is required. The air is cooled and dried. Humidity in the air source will cause condensation and freezing which will cause the pellets to stick together. In the Cold Jet cleaning system, the CO_2 pellets were mixed with this driver air at the pelletizer. The compressed air forces the pellets down a hose through the nozzle where they are sprayed on the test article. The Alpheus system propels the pellets to the nozzle at 40 psig and then mixes the pellets with the compressed air at the nozzle.

The motion of the nozzle was controlled by the end effector on a T3-776 Cincinnati Milacron robot with 6 degrees of freedom. The positional repeatability of the robot arm is ± 0.1 inches. Robot motion patterns could be saved and reloaded into memory later. Although the robot is equipped with a teach pendant, a graphical dynamic software system allowed for computer simulation of the hardware setup and offline programming. In this way the robot motion could be programmed and checked for interferences, processing time durations, and test article coverage without actually risking an unintended motion which might cause some damage to the robot or the test article. The robot was mounted on a track base and translation table and is equipped with a controller. This provided the robot with the additional freedom of lateral motion. The translation table controller was interfaced with the robot controller. The positional repeatability of the translation table was ± 0.2 inches. The robotic set up is shown in figure 2.

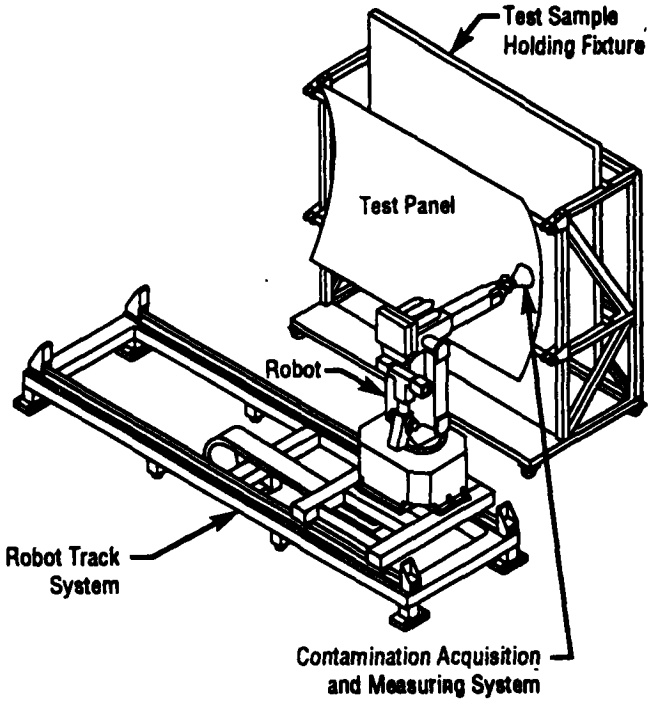


Figure 2 Robotic Setup

The robot and test article were housed in a portable class 100,000 clean room. A high volume contamination acquisition and filtering system was used to remove as much of the rebounding CO_2 gas with contamination from the area as possible. An accumulator housing was installed surrounding the nozzle to directly pull the gas away from the area. This system removed contamination from the clean room to minimize any redeposition of the contamination and was also connected to an experimental real time contamination monitoring system.

The real time contamination monitoring system consisted of a thermal quartz crystal micro-balance for measuring molecular contamination and a laser particle counter instrumentation system for measuring particulate contamination. Results from the real time analysis systems would be compared to standard analysis procedures obtained by conventional methods performed in chemical analysis laboratories at MSFC, Martin Marietta laboratories at Manned Space Systems (Michoud) in New Orleans, Martin Marietta laboratories at Astronautics Group in Denver, South East Analysis Services Laboratories in Huntsville, Al, and Lockheed Laboratories in Sunnyvale, California.

Cleaning Process

The cleaning process has been recorded with high speed photographic equipment in order to observe the phenomena of the pellet impact on the substrate. It has not been possible to precisely determine the exact mechanics of the impact, but the cleaning effect appears to be due to a multimode transformation of the pellets. Contributing to the cleaning is the mechanical impact of the pellet as it strikes the substrate, a thermal shock to the contamination layer which embrittles the contamination making it easier to remove, sublimation of the pellet which results in a large volume expansion which blows the contamination away, and a possible momentary liquid phase which acts as a solvent. System parameters which influence the effectiveness of the cleaning process are clean room environment, humidity, nozzle translation speed, nozzle angle, system pressure and pellet mass flow rate, nozzle stand off distance, pellet density and size, the number of passes over the same area, the percent spray overlap per pass, contaminant being removed, and surface finish of the material being cleaned.

Cleaning Dynamics

Conventional grit blasting penetrates the contamination layer and actually removes a small quantity of substrate during the impact. The CO_2 pellet penetrates the contamination layer and sublimes on impact with the substrate. This phase change of the pellet dissipates the impact energy and causes no damage to the substrate. Very thin material may be blasted with the CO_2 pellets without causing any peening. The large volume change of the CO_2 during the sublimation helps to lift and blow the contamination from the surface resulting in cleaning a larger area per impact than grit. Direct contact with a pellet chills the neighboring area of contamination and weakens the bond to the substrate making the bond more fragile and easier to break. At very high pressures and mass flow rates, the pellets start to fracture as they near the surface, resulting in an increase in the number of impacts on the surface.

Wastes

All cleaning processes obviously result in waste from the contamination itself. Most process also have a secondary waste from the cleaning agent. High pressure water spray will result in a large quantity of water which will require treatment and disposal of a concentrated sludge. Aqueous solutions although not harmful to the environment in themselves must be processed because of the entrained contamination cleaned from the substrate. Chemical solvents must be collected and stored as hazardous waste. Any hand wipes or cloths must also be handled as hazardous waste. The CO_2 cleaning system results in no secondary waste since the CO_2 pellets sublime and return to the atmosphere. The CO_2 used in the process is obtained from the waste gas stream from various other manufacturing processes and would have been exhausted to the atmosphere anyway.

Test Setup

Test panel configurations largely consisted of large flat plates, small 2" X 2" flatwise tensile test blocks, and complex geometries such as the isogrid pattern on the Titan IV payload fairing. Test panels were cleaned prior to the test with a conventional cleaning method and tested to assure that the contamination on the surface of the panel was less than 1 mg/ft^2. Known amounts of various contaminants were then applied to the surface of the panel. The panel was cleaned using the CO_2 cleaning system and the panel was again checked for cleanliness. Analysis utilized various standard techniques for determining non-volatile residue (NVR) and particles such as manual solvent flushes, tape lifts, optical scanning methods, and cloth wipes.

RESULTS

Titan IV Payload Fairing

Titan IV cleanliness requirements vary depending upon the payload being used. The most stringent requirements would be for a molecular contamination level of less than 1 mg/ft^2 and a particulate level under Visibly Clean Level 2. Visibly Clean Level 2 requires that the surface be free of all particles and oil films when observed by the unaided eye at a distance of 12-24 inches with 100-125 footcandles light intensity at the surface.

Tests performed were divided into three groups. The first group of tests was designed to identify the effectiveness of five different spray patterns in removing each of three different contaminants plus a mixture of all three contaminants. The second group of tests was designed to determine the number of times each spray pattern must be repeated in order to clean the test panel to the required degree of cleanliness. The third test was a validation of the most promising method while cleaning an actual ⅛ length section of a payload fairing.

The CO_2 cleaning system was proven to be an effect process for the cleaning of the payload fairing for the Titan IV rocket. It was determined from the first group of tests that a vertical nozzle motion was the most effective spray pattern. The second set of tests shows that most of the contamination is removed on the first pass since the upper layers of contamination are bonded to the lower layers and not to the metal surface itself. The

most effective nozzle orientation was proven to be one in which the nozzle is aimed in the direction of motion. This method blows the rebounding gas ahead of the nozzle and away from the previously cleaned section of panel.

The validation test utilized the most effective of the five spray patterns tested. The pattern consisted of a four pass vertical nozzle motion with a 50% spray overlap. The first pass held the nozzle perpendicular to the surface. The nozzle direction lagged its motion by 30° during the second pass and lead its motion by 30° during the third pass. These passes aid in cleaning around and behind obstructions. The fourth pass repeated the first perpendicular pattern. After the entire panel was cleaned, this entire process was repeated once more. The results show that the CO_2 cleaning system is capable of meeting the Titan IV requirements. Due to the geometry of the isogrid, some areas of the panel were obstructed by 4" deep stringers. Final NVR values averaged below 0.2 mg/ft^2 in both obstructed and unobstructed areas of the isogrid.

The experimental real time contamination monitoring system data showed the approximate trends in cleanliness level as the standard NVR Freon rinse tests but there was insufficient data to correlate the results and develop an equation.

Shuttle External Tank

CO_2 cleaning for use in preparing the shuttle external tank for ablator bonding was evaluated by bonding an ablative material to materials cleaned by either the CO_2 system or conventional chemicals. Various adhesion tests were conducted to compare the effectiveness of the CO_2 cleaning system to the standard process. The difference between the adhesive strength of the CO_2 cleaning system varied within ±10% of the adhesive strength of the panel cleaned by chemical processes.

Biological Film Removal

Microscopic biological organisms grown on metallic test slides have proven difficult to remove due to their tight bond with the metallic surface. In the past, slide have either been soaked in acid, and scrubbed and scraped with limited success, or simply discarded. Multiple passes of high intensity CO_2 pellet cleaning was able to remove the film from these slides to a much greater degree than any other method attempted. After cleaning the slides were etched so that the organisms would show luminously under a microscope. Clean and unclean areas of the slide could be compared side by side. The clean side showed almost no luminous areas.

Additional Testing

The system was demonstrated and tested on many other articles. It was shown that the CO_2 system is capable of stripping coatings as well as performing cleaning. Dry film lubricants were easily removed from nut bearings. Various thermal protective coatings and ablators were striped from aluminum test panels. By using special nozzles which fracture the pellets as they exit the nozzle, printed circuit boards were cleaned to remove the flux left on the surface after manufacturing.

Both Cleanliness and bonding test were conducted on ASRM candidate materials. 2" X 2" painted steel plates were cleaned with conventional cleaning agents and the CO_2 system. Test blocks were bonded to the cleaned surface which were then subjected to a flatwise tensile load. In all tests the failure occurred in other areas other than the cleaned surface bond indicating the CO_2 system is as effective as conventional cleaning procedures.

Various other steel, aluminum, and carbon phenolic materials were also cleaned. In all cases final NVR readings were in the low mg/ft^2 reading. It was determined that the contamination evacuation system and clean room environment were important factors in preventing the redeposition of contaminants after cleaning. During long blasting operations, the CO_2 pellet temperature could lower the test article temperature sufficiently to cause condensation. This condensation could result in flash corrosion. As the water droplets attach to the substrate they also bring with them other contaminants which are left on the surface after the water evaporates. It is

important that this condensation be eliminated to ensure a high level of cleanliness is maintained. New generation CO_2 cleaning systems are designed to slightly reheat and dry the test article in order to prevent the formation of condensation.

CONCLUSION

CO_2 cleaning systems are an effective cleaning system for the removal of many types of contamination on a wide variety of materials. The system is safe for use on even extremely thin materials, not hazardous, and not toxic. It is capable of meeting the goal of reducing the CFC usage per EPA requirements. The system is capable of quickly removing gross contamination and can be fine tuned to achieve a high level of cleanliness. Parameter optimization is required for each geometry although general patterns exist to provide a good starting point for any testing. With proper contamination evacuation, humidity control, and clean room environment it is possible to achieve NVR levels below 0.1 mg/ft^2.

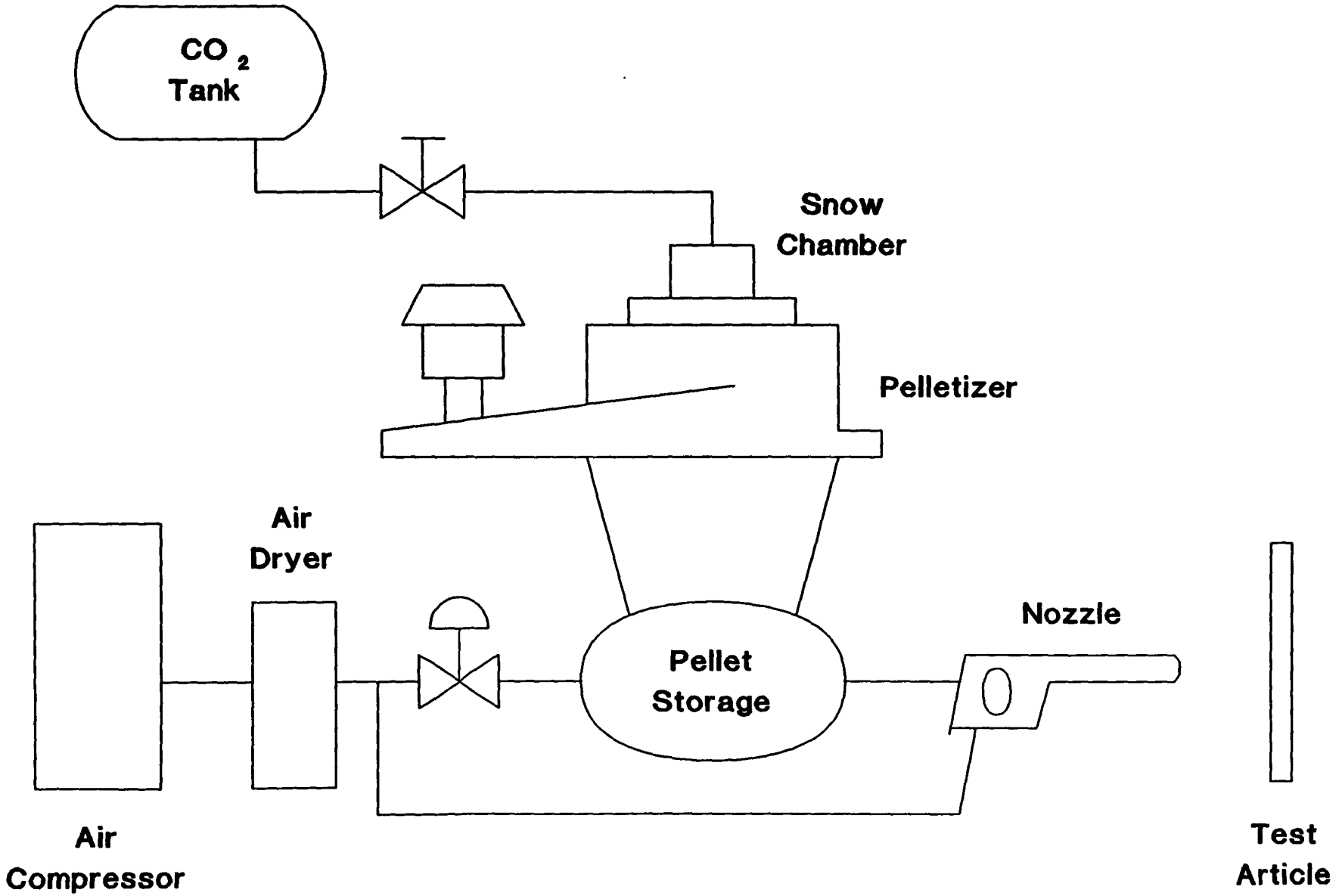


Figure 1: CO_2 Cleaning System Setup

MATERIALS SCIENCE

(Session E5/Room A2)

Thursday December 5, 1991

- **Applications of Biologically-Derived Microstructures**
- **Structural Modification of Polysaccharides: A Biochemical/Genetic Approach**
- **Cryogenic Focusing, Ohmically Heated On-Column Trap**
- **Study of the Effect of Hydrocarbon Contamination on PFTE Exposed to Atomic Oxygen**

PRECEDING PAGE BLANK NOT FILMED

N92-22728

ADVANCED COMPOSITE APPLICATIONS FOR SUB-MICRON BIOLOGICALLY DERIVED MICROSTRUCTURES

J. M. Schnur, R.R. Price, P.E. Schoen
Center for Bio/Molecular Science and Engineering
Code 6090, Naval Research Laboratory
Washington, D.C. 20375-5000

Joseph Bonanventura
Director, Marine Biomedical Center
Duke University Marine Laboratory
Pivers Island
Beaufort, North Carolina

Douglas Kirkpatrick
Scientist
Scientific Applications International Co.
McLean, Virginia

ABSTRACT

Advanced materials of the 21st century require development of new materials with significantly improved properties. These new materials may lead to solutions of existing problems, and additionally open the door to entirely new technologies. A major thrust of advanced material development has been the area of self assembled ultra-fine particulate based composites (micro composites). We report on the application of biologically derived, self-assembled microstructures to form advanced composite materials. Hollow 0.5 micron diameter cylindrical shaped microcylinders self assemble from diacetylenic lipids. These microstructures have a multiplicity of potential applications in the material sciences. Exploratory development is proceeding in application areas such as controlled release for drug delivery, wound repair, and biofouling as well as composites for electronic and magnetic applications, and high power microwave cathodes.

INTRODUCTION

Recent advances in self assembly and nanotechnology suggest a number of difficult material problems faced by the engineers of the 20th century are likely to be solved early in the 21st. Opportunities abound for new products based upon "smart" materials or materials with significantly improved properties.

NRL's Center for Bio/Molecular Science and Engineering (CBSE) has focused its initial efforts in self assembly and bio/molecular materials on fabrication of submicron structures. Much of our effort has been aimed at modification of lipids to enable the formation of technologically interesting microstructures. Our ability to fabricate these structures represents an exciting synthesis of biotechnology and molecular engineering.

This paper will focus on one particular type of self assembled micro structure that was discovered at NRL. By inserting diacetylene groups into phosphocholine lipids, we have been able to form sub micron diameter hollow cylinders called tubules. These cylinders are very much like micron sized soda straws with outer diameters from 0.05 microns to 0.7 microns or so and lengths from 10 microns to over 1 millimeter[1,2,3,4,5,6] (See Figure 1).

PRECEDING PAGE BLANK NOT FILMED

BACKGROUND

Lipids are the basic building block of biological membranes. The bilayers of lipid self assemble to form either two dimensional sheets, or liposomes which are spherical structures composed of a lipid bilayer enclosing an aqueous space. A number of important applications, most involving targeted release, have been identified for liposomes. Our laboratory has been one of the pioneering centers in the development of an artificial blood surrogate utilizing liposomes as the encapsulant for hemoglobin. While the spherical geometry has a number of important benefits, other geometries may also be of interest. For instance, a hollow cylinder can provide a narrow channel for diffusion leading to zero order diffusion rates (i.e. a constant rate of release that is independent of time) for controlled release applications. A suitably conducting hollow cylinder provides opportunities for the development of advanced high dielectric materials[7,8,9,10]. Other geometries may such as template formed tubule channels or euctectic devices also have their own particular technological niche.

CONTROLLED RELEASE

Since man first set sail on the oceans of the world, sailors have been plagued by biofouling and deterioration of their vessels caused by both plant and animal species. One early solution was copper sheathing chosen to protect wooden hulls against the ravages of the teredo worm. Today the modern equivalent of copper sheathing is a polymeric ablative or self polishing paint, using copper powder or cuprous oxide as the primary toxicant. Although copper as an antifoulant has stood the test of time, it is often more effective against animal species than plant growth. In order to improve on the performance of copper the addition of many other metal species have been employed, such as mercury, arsenic, cadmium, lead and tin. Currently, in more enlightened times, these heavy metals have been abandoned as persistent toxicants that have adverse effect on the environment and more directly on man's health as a consumer of fish and shellfish.

In addition to environmental concerns, as fuels become more expensive and fossil fuel reserves begin to shrink the economic costs associated with biofouling begin to rise. Marine biofouling growth on underwater ship hulls increases hydrodynamic drag and hull weight due to the increasing biomass. If reasonable service speeds for the ship are to be maintained, power output must be increased, resulting in higher consumption of fuel and increased wear on the machinery. Increases in fuel consumption exceeding 10% are common, as are decreases in operational hull speeds of up to 16%.

In an effort to ease the economic pressures that continue to mount on governments, industry and individuals, service cycles are being increased. The U.S. Navy has lengthened operational cycles to 5 years, with 7-10 year cycles under consideration to further reduce maintenance costs. When antifouling paints fail early on in the service cycle, it may not be possible to haul and paint the ship ahead of schedule due to lack of funds or facilities availability. This situation would result in higher fuel costs and decreased performance associated with a fouled hull for a greater portion of the maintenance cycle, but would have to be tolerated unless extended antifouling performance can be attained.

In order to achieve the goal of less polluting yet effective antifouling paints, strict control of the release rates in copper based paints in addition to the entrapment and controlled release of alternate compounds must be accomplished. These alternate compounds are often active at levels far lower than those needed for copper. Thus proper control of release rates is necessary to prevent these alternate compounds from being discharged from the coating in excess of actual requirements. Conservative, i.e. low release rates led directly to long service lifetimes for the paints.

Not only must an antifouling paint offer performance in service; it must meet increasing demands from governmental regulators concerning water pollution standards (< 18 ug/l maximum), air quality standards for volatile organic solvents (VOC) and the occupational and health regulations governing the application and disposal of antifouling coatings.

The small pore size and long cylindrical shape of the tubule offer significant opportunities for control of long term control of release. By controlling the viscosity of the polymer, the length and diameter of the tubule, and the permeability of the paint matrix, variations of the release rate of several orders of magnitude have been demonstrated. The tubules isolate the encapsulant from the environment thus providing a mechanism for enhanced chemical stability of the encapsulants. The tubules are quite compatible with existing paints and may even offer some improvements to the ultimate mechanical properties of the composite [11,12,13,14].

Methods for Environmental Exposure Testing:

The metallic microcylinders utilized in the study averaged 0.5 microns in diameter and ranged from 10-250 microns in length with interior diameters ranging from 0.25-0.4 microns. Once dry the microcylinders act as microcapillary tubes entrapping and retaining a range of liquid materials. Release rates are linear and dependent on the encapsulant and the molecular weight and cross linking of the carrier vehicle.

Twenty, 11 cm fiberglass rods, which had been cut from 0.35 mm diameter stock, were coated by dipping with the desired formulation of antifouling coating. Following air drying for at least 48 hr., the rods were mounted in a floating array consisting of a rectangular float of PVC pipe to which a diamond polypropylene fish impound netting was attached (Figure.2). The rods were attached to the net with rubber grommets and the entire array was attached to a raft in the field at Coconut Island, Hawaii, or Taylors Creek, Beaufort, North Carolina. At the first sign of fouling rods were withdrawn and examined to determine the composition and relative percentage cover of fouling organisms.

Tetracycline was used in the initial studies of release rates from coatings and microstructures as it is easily quantified by spectrophotometric analysis in water. In addition tetracycline is a registered antifouling agent and therefore was considered safe in this application. Initial findings indicated tetracycline could be released at linear rates from both epoxy resin films, and from vinyl based paints. It is interesting to note in Figure 3 that when not encapsulated the tetracycline is found to release in a matter of a few hours from VYHH coatings; however, when in an encapsulated sample, tetracycline continues to release after 500 days of use.

In order to explore the possibility of lowering the need for antibiotics or persistent copper toxicants, two further approaches were tested. First was the use of isothiazolone which is an experimental antifouling agent which has been shown to be non-persistent in the marine environment (Harrington, 1989). When encapsulated in both copper and iron microcylinders and added to a vinyl matrix this additive was shown to be effective at repelling fouling marine species in testing (Figure 4).

One other approach to the development of non-toxic or biodegradable coatings is the use of extracts from the Sea Pansy (Renilla reniformis), and structural analogs of these compounds. Figure 5 illustrates a pair of test rods which were taken from a sample exposed at Coconut Island, Hawaii for six months. It can be clearly seen that the experimental rod has successfully repelled fouling during the test period with concentrations of 2% by weight active agent.

Figure 6 is a comparison of the fouling characteristics of four coatings which consisted of a control baseline self polishing coating consisting of 1) methacrylate paint base on the test rods, 2) copper microtubules in the paint base, 3) encapsulated isothiazolone, and 4) encapsulated isothiazolone and a crude extract of renilla. The paint base is observed to become heavily covered by fouling (both micro and macro fouling are considered on an equal basis) by day 20, after which the fouling is reduced somewhat. This may be a result of predation of the fouling community during testing, as the coverage is again observed to increase following 90 days exposure to greater than 90% of the test surface. In the case of the paint base containing 5% by weight copper tubules the fouling is observed to increase to 50% coverage at day 40 and then is reduced to levels less than 40% until day 120. Again this may be due to the effect of predation or the settlement rates of larva as the test progressed. Test paint with encapsulated isothiazolone was observed to offer more consistent performance as compared to paint containing only copper tubules. Maximum coverage for all fouling species was less than 40% at day 40 and was observed to decrease during the period of testing. In contrast the addition of renilla extract to this mixture

provided the best initial performance of the formulations tested; however, by day 120 the average settlement had exceeded that of the isothiazolone only. As the loading of microtubules remained constant by weight, the isothiazolone in this last sample was reduced by half, and thus it may be that the renilla which was encapsulated had either become depleted or become ineffective against the fouling species settling at the end of the testing period.

Controlled Release: Summary

A number of major issues must be resolved before the ultimate utility of this approach can be determined. The variable controlling release rate must be quantified and optimized for each particular application. Important variables are the molecular weight of the antifouling agent, diameter and length of the tubules, permeability of the incorporating polymer as well as the matrix paint. While we have been able to make gallon quantities of test paints in our laboratory, the question of scale up has still to be seriously addressed. Cost is always an important issue. We had thought that lipid costs (up to $4000 per pound) would be a serious problem for technology transfer. However, we have had a recent success at recovering the lipid for metalized tubules and recycling that lipid to make more tubules which were subsequently metalized.[1] In addition, as the commercial production of this class of phospholipid compounds is undertaken the economies of scale should (as with other high technology chemical systems such as teflon) be reduced to a reasonable level. Further reductions of cost will be realized with large batch processing which would reduce labor costs, thus production costs of microstructures should not be barrier to use. Costs of the metal, biofouling control agent, and matrix, coupled with the processing costs will determine the commercial viability of this approach.

The system holds promise as a means of providing a controlled release mechanism for antifouling paints with significant advantages over existing formulations. The inclusion of microencapsulated toxicants is effective in short term assays designed for the study of larval settlement, and appears to be a means of providing controlled release independent of the paint vehicle.

In addition, the use of microcylinders instead of rosin allows for a more robust coating which is better able to resist erosion in high flow areas, and which will be better able to survive long periods of immersion in seawater.

While antifouling has been our first attempt at utilizing tubules for controlled release, several other possible applications are currently being evaluated. These include wound healing, enzyme storage and delivery, and long term antimicrobial and antifungal coatings. As a controlled delivery system microcylinders offer advantages in antimicrobial control systems for metal working fluids and coolants, prevention of deterioration in paper processing plants, and stabilization of kerosene and diesel fuels in storage.

ELECTRONIC AND MAGNETIC MATERIALS

When we first began to assess the potential applications for these micron sized "soda straws", the physicists in our research team were struck with the possibilities for enhanced electronic material applications. This is due to their large and adjustable aspect ratio, their cylindrical geometry, and very thin walls of the tubule.

Experiments and detailed calculations on the interaction of electromagnetic radiation with conducting hollow cylinders in composites have shown significant improvements[7,8] in temperature and frequency stability of the dielectric constant, both real and imaginary, as well possibilities for exploiting their highly anisotropic properties.[3,4]

1

In order for these calculations to be confirmed and the potential applications assessed a process for rendering the tubules conductive had to be developed. This process[5] has now been refined so a coating can be placed on the inside and outside of the tubule. The thickness of this coating can be varied within a 150Å tolerance limit within the thickness range of 200Å to 1200Å. Copper was the first metal to be electrolessly deposited onto the lipid tubules, but now they are coated with nickel, nickel-iron-boron (a permalloy), gold, or palladium. The direct current conductivity of our current coatings appears to be about 2-3 orders of magnitude below the values expected for the bulk metal. Improving the metallization process is currently a subject of intense interest in our laboratory. Nevertheless the current coatings have demonstrated that tubule based templates offer significant advantages for the development of improved electronic materials. Behroozi et al[7] have reported that at 10% volume loading of permalloy coated tubules in an epoxy matrix a real dielectric constant approximately equal to 50 along the alignment axis at a frequency of 9.5 Ghz has been achieved for tubules of average length of 30 microns with a lipid diameter of 0.5 micron and a 1000Å thick metal coating. This result agrees with their electrodynamic calculations. The paper suggests that far higher values may be achievable with longer tubules and improved metal coatings. Such results suggest the possibility for applications such as high density packaging for microelectronic applications, high dielectric materials for miniaturized microwave device applications, and phased arrays and waveguides for radar applications. Tubules may also serve as templates for fabricating field emitting arrays for microwave cathode applications. For any of these applications to be introduced into actual use the electroless metal deposition process will have to be improved and appropriate matrix materials identified, e.g. ceramics or polymers etc.

Field Emission Applications

Recently vacuum field emission was realized from a cathode fabricated of metallized microtubules. The surface micro-morphology of sharp cylindrical emitting tips is produced by using tubules, as the template for metal deposition, and subsequently forming an aligned composite of these metal microstructures[10].

Limitations of present electron source technology are manifested in microwave devices high energy particle accelerators, laser pumping, and other fields which utilize electron beams as a means of energy transfer. Presently available electron sources are divided into three categories: thermionic emitters, laser-activated photo-emitters and field emitters. Included in this last category of field emitters are both exploding field emitters, sometimes termed plasma cathodes, and vacuum field emitters which do not form an intermediate plasma. The structures we discuss here may provide a means to achieve electron beam brightness in excess of 10^6A/cm2-rad2 from an unsaturated field emission cathode, by using the electrostatic lensing produced near the tip of the hollow emitter micro structures. The generation of macroscopic electron beam currents through vacuum field emission from a large number of emission sites requires a surface with a complex micro structure . To date, fabrication of surfaces suitable to this task has been dominated by micro lithographic techniques. In these processes, masks are used in conjunction with etching or deposition techniques to produce arrays of micron scale cones or wedges recessed only a few microns from apertures in a gate structure[15,16]. The micro structure composite cathode materials described here do not use such an electrode configuration with gate, and are similar in this respect to more conventional electron source materials. The structure of an array of hollow cylinders protruding a uniform distance from a base electrode is preferred over an array of pyramids or cones because of the larger available emission area. As additional advantages, these materials appear to be relatively insensitive to the background vacuum pressure, and operate at DC.

The necessary local enhancement of applied electric field is produced by the geometry of the exposed tubules: their height, inner and outer radius, the average spacing between nearest neighbors, the radius of curvature of the metal wall at the edge of the exposed hollow cylinder, and the character of the surface in the vicinity of the exposed edge. Detailed numerical simulations of the electrostatic field in the vicinity of the hollow cylindrical structure have shown that field enhancement factors in the range of 150--250 are readily achieved with the 0.5 micron diameter tubules protruding a height of 10 to 15 microns above the base surface. The intrinsic surface roughness of electrolessly deposited metal film that makes up the outer tubule surface would probably increases this nominal enhancement factor by an additional factor of 2-4, yielding an expected range in the enhancement factor of B= 300--1000.

The hollow nature of the protruding tubule micro structures (Figure 7) that make up field emitter arrays is of particular interest, because it provides an electrostatic lensing effect for the emitted electrons. Strength of this electrostatic lensing depends on the inner and outer radii of the tubule, and the position of the emission area on the end of the tubule structure. The thinner the metallic wall, the greater is the self focussing effect of the structure, and the more collimated the emitted micro beamlet becomes. For suitably fabricated structures, with thin wall thickness near the emission tip, the previously cited numerical simulations have indicated that normalized electron beam brightness well in excess of 10^9 A/cm2-rad2 can be achieved[17].

In conclusion, we have described measurements of vacuum field emission from an unsaturated field emitter array fabricated from a composite of self-assembling biomolecular micro structures. Micro structure composite materials offer an interesting alternative to micro lithographic techniques for the achievement of complex surface micromorphologies. Complex biological systems and organic molecules, in particular self-assembling biomolecular micro structures, offer a wide variety of micro structure geometries potentially useful for application in physical systems. The hollow, thin walled , high-aspect ratio tubule micro structures discussed here might provide a surface micro morphology suited to the generation of high current, high brightness electron beams. An identical structure would be difficult to generate using existing micro lithographic techniques.

SUMMARY AND CONCLUSION

The ability to on a molecular level engineer sub micron particles represents a seminal advance in the development of advanced composite materials leading to a number of new composite materials with significantly enhanced mechanical and electronic properties. We believe this to be true as we have substituted natures ability to self-assemble complex molecular structures rather than to rely on often expensive and sensitive technological equipment to achieve the same result.

We are now actively pursuing applications for the lipid tubule microstructures. These applications range from antifouling to the development of advanced electronic materials. Antifouling controlled release paint systems have been successfully developed and tested. Other applications such as wound healing, antifungal paints, and bioremdiation systems are currently under evaluation. Prototype fabrication of tubule based high dielectric and field emitting materials suggest possible commercial applications for high resolution displays and advanced microwave devices. It is clear that advanced materials research and development is entering a new stage due to recent advances in "self assembly" and possibility of the design of molecules for the "engineering" of supermolecular micron sized structures for specific applications.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Naval Research Laboratory, Office of Naval Research, Office of Naval Technology, Defense Advanced Research Projects Agency, and the National Science Foundation for partial support of the research described above. This work could not have done without the work of the entire Center for Bio/Molecular Science and Engineering tubule project research team and their contributions are gratefully acknowledged. We also wish to thank Dr. Bhakta Rath, Dr. William Tolles, Dr. Dick Rein, Dr. Ira Skurnick, and Captain Steve Snyder for many helpful discussions.

REFERENCES

1. P. Yager and P. Schoen (1984) Mol. Cryst. Liq. Cryst. 106, 371-381.

2. J.H. Georger, A. Singh, R.R. Price, J. Schnur, P. Yager and P. Schoen (1986) J.A.C.S. 109,169-6175.

3. Schoen, P.E., Yager, P., Schnur, J.M.; (1991) U.S. Patent #4,990,291.

4. J.M. Schnur, P.E. Schoen, P. Yager, R. Price, A. Singh, J.H. Georger,(1989), U.S. Patent #4,877,501.

5. J. M. Schnur, A. Singh, (1989), U.S. Patent #4,867,917.

6. D. Chapman, (1982), U.S. Patent #4,348,329

7. F. Behroozi, M. Orman, R. Reese, W. Stockton, J. Calvert, F. Rachford and P. Schoen (1990) J. Appl. Phys. 68, 3688-3693.

8. W. Stockton, J. Lodge, F. Rachford, M. Orman, F. Falco and P. Schoen (1991) J. App. Phys. (in press).

9. J. M. Schnur, P.E. Schoen, P. Yager, R. Price, J. M. Calvert, and J. H. Georger, (1990), U.S. Patent # 4,911,981

10. D.A. Kirkpatrick, J. M. Schnur, P.E. Schoen, W. M. Manheimer, U.S. Patent (Allowed 30 July 1991), Serial #07/589757

11. R. Price and M. Patchan, (1991) J. of Microencapsulation,8, No 2.

12. R. Price, M. Patchan, B. Gaber, (1990), Proceedings, 7th International Symposium on Microencapsulation, Glasgow, Scotland.

13. R. Price, M. Patchan, D. Rittschoff, A. Clare, and J. Bonaventura (1991) Transactions of the Institute of Marine Engineers, Inst. of Marine Eng. Press, Rhian Bfuton, Ed.

14. R. Price and R. Brady, U.S. Patent (Allowed 23 April 1991), Serial #07/343762

15. C.A. Spindt, I. Brodie, L. Humphrey and E.R. Westerberg (1976) J. Appl. Phys. 47, 5248.

16. H.F. Gray, G.J. Campisi and R.F. Greene (1989) Proc. IEDM 89 (Washington, DC) 776.

17. D. Kirkpatrick, P.E. Schoen, W. Stockton, R. Price, S. Baral, B. Kahn, J.M. Schnur, M. Levinson, B.M. Ditchek, (1991) IEEE Transactions Sci, Special Issue on Vac. Discharge Plasmas.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

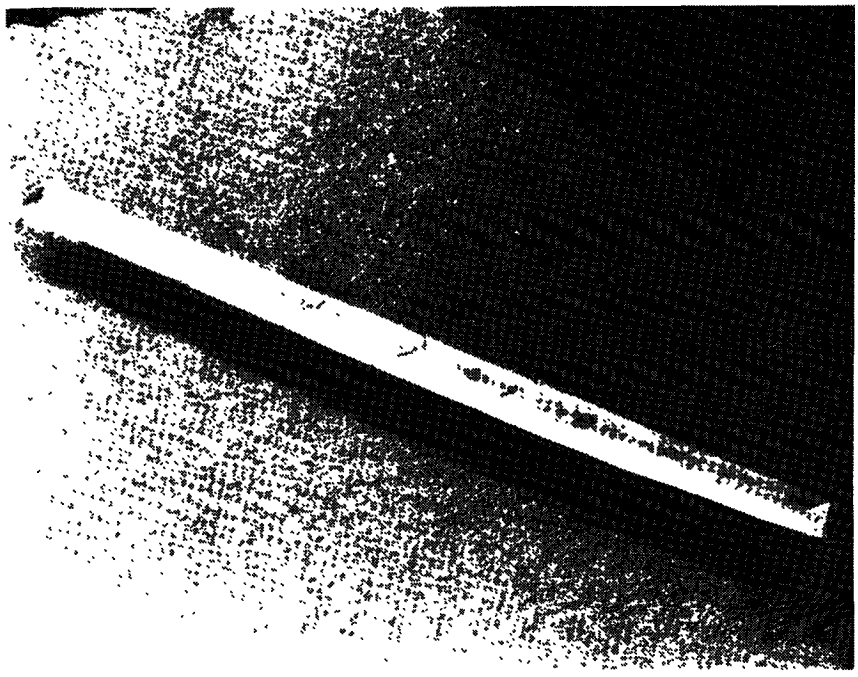


Figure 1 Electron micrograph of a typical lipid tubule.

Figure 2 Photograph of rods were mounted in a floating array.

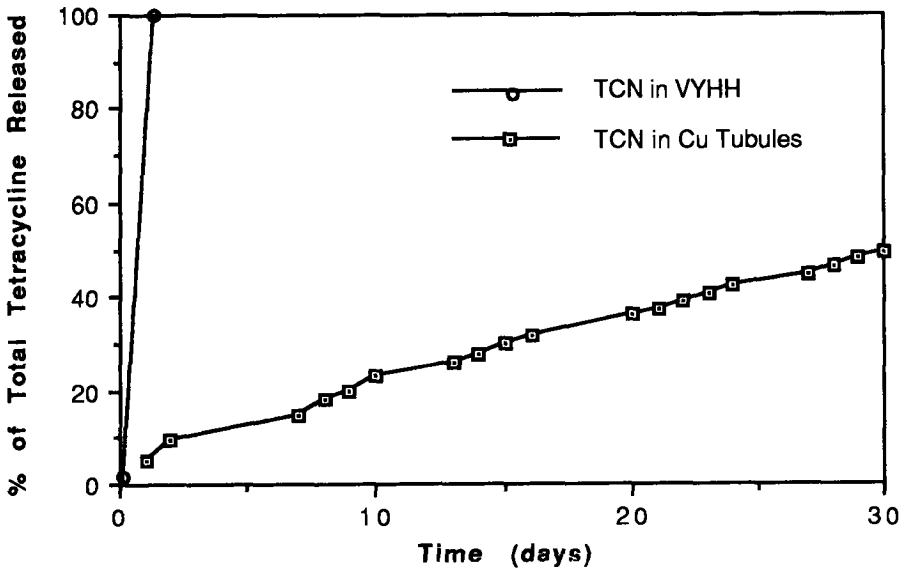


Figure 3 Encapsulated vs. non-encapsulated release rates for tetracycline in a vinyl matrix.

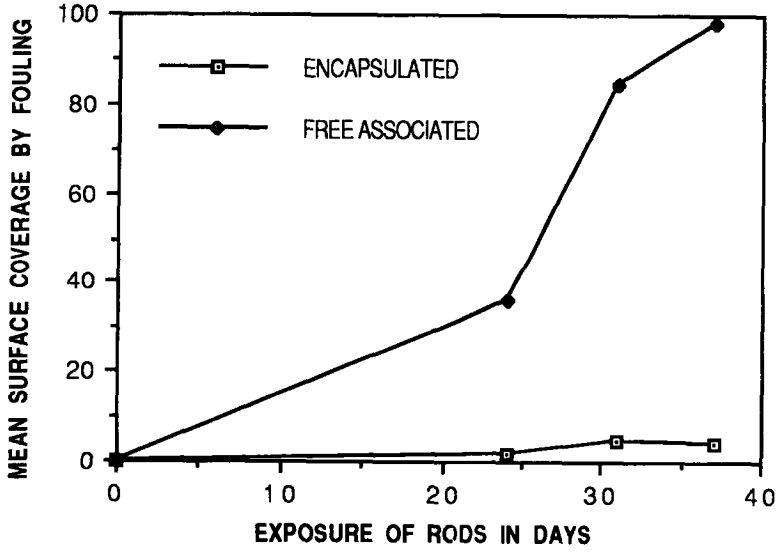


Figure 4 Comparison of effectiveness of encapsulated vs. non-encapsulated isothiazolone on marine fouling.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

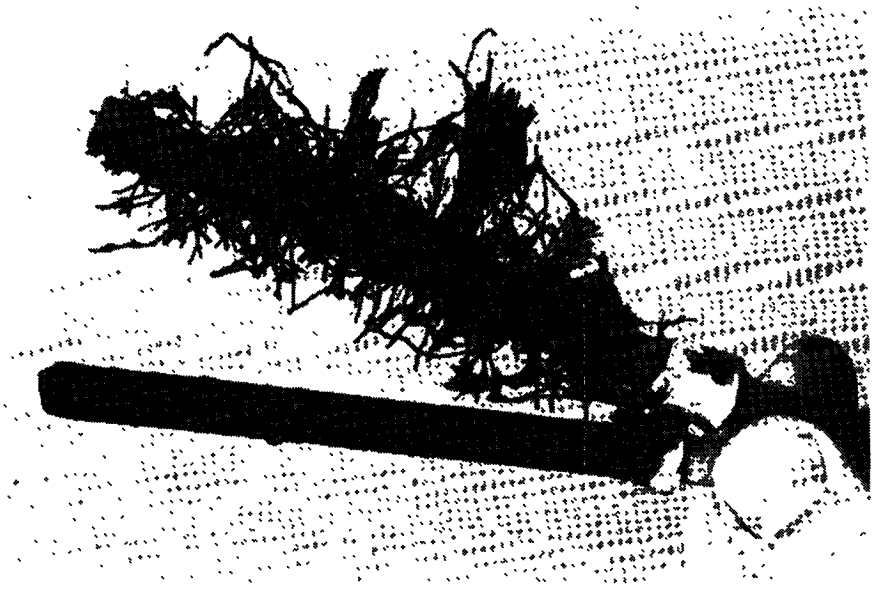


Figure 5 Test rods from Beaufort, N.C. effectiveness test, illustrates effectiveness of microencapsulated natural product.

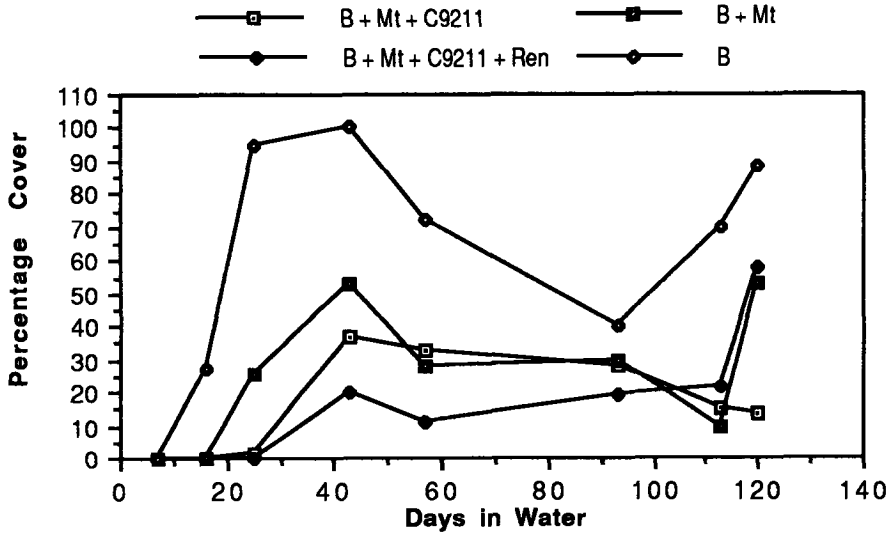


Figure 6 Comparison of fouling characteristics in four coatings consisting of 1) copolymer methacrylate paint, 2) Copper microtubules in paint, 3) Isothiazolone encapsulted, 4) isothiazolone and extract of renilla.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Ni tubules in epoxy, plasma etched. SEM

Figure 7 **Electron micrograph of a tubule based cathode array.**

N92-22729

STRUCTURAL MODIFICATION OF POLYSACCHARIDES: A BIOCHEMICAL-GENETIC APPROACH

Roger G. Kern and Gene R. Petersen
California Institute of Technology
Jet Propulsion Laboratory
4800 Oak Grove Dr (89-2)
Pasadena, CA, 91109

ABSTRACT

Polysaccharides have a wide and expanding range of industrial and biomedical applications. An industry trend is underway towards the increased use of bacteria to produce polysaccharides. Long term goals of this work are the adaptation and enhancement of saccharide properties for electronic and optic applications. In this report we illustrate the application of enzyme-bearing bacteriophage on strains of the enteric bacterium *Klebsiella pneumoniae*, which produces a polysaccharide with the relatively rare rheological property of drag-reduction. This has resulted in the production of new polysaccharides with enhanced rheological properties. Our laboratory is developing techniques for processing and structurally modifying bacterial polysaccharides and oligosaccharides which comprise their basic polymeric repeat units. Our research has focused on bacteriophage which produce specific polysaccharide degrading enzymes. This has lead to the development of enzymes generated by bacteriophage as tools for polysaccharide modification and purification. These enzymes have been used to efficiently convert the native material to uniform-sized high molecular weight polymers, or alternatively into high-purity oligosaccharides. Enzyme-bearing bacteriophage also serve as genetic selection tools for bacteria that produce new families of polysaccharides with modified structures.

INTRODUCTION

Polysaccharides are a ubiquitous, integral, and often invisible part of the manufacture and production of a very broad spectrum of products. Traditional uses of these materials include, but are not limited to, food additives, pharmaceutical chemicals, oil well drilling additives, industrial coatings, industrial cleaners, explosives, paint additives, paper chemicals, printing chemicals, photographic chemicals, textile chemicals, and additives for ceramics and metals[1]. Recent industrial applications of polysaccharides include drag-reducing agents for ships and water jet-cutting applications, photo-resist applications in the manufacture of integrated circuits, and the synthesis of new composite materials. Promising biomedical applications of polysaccharide themselves include a role in drug delivery and wound treatment[2].

Synthesis of new oligosaccharides with biological activity is an area of intense and growing interest. The discovery of the important and diverse role of saccharides in biological processes has lead to an increased demand for practical routes to gram scale quantities of saccharide-based compounds. Considerable progress has been made towards enzymatic approaches to oligosaccharide synthesis by a number of ingenious methods. Despite efforts in towards general methodologies, saccharide synthesis has been one of the most challenging areas of biochemistry[3]. The wide diversity of oligosaccharides produced by enzymatic polysaccharide degradation to its basic repeat unit provides a potential source of starting material for glycosylated pharmaceutical agents.

Hundreds of naturally occurring bacterial polysaccharides provide an untapped resource of chemically, structurally and functionally diverse biopolymers. Although polysaccharides have traditionally been derived from

plant and animal origin, the industry trend has been toward replacing these sources with microbially produced polymers. This is due to: (a) economic concerns, such as supply reliability and long term cost control, (b) quality control issues, such as product purity, and (c) application issues, which require consistency and specificity of the chemical and physical properties of the polysaccharide.

Despite these advantages, there are still technological problems which limit the production and isolation of polysaccharides from bacteria. Microbial bioreactors that produce polysaccharides generate a viscous fermentation liqueur that interferes with oxygen transfer, which in turn, limits microbial growth, and ultimately reduces polymer yield[4]. In addition, separating large amounts of polysaccharide from cells by conventional centrifugation is difficult because the polysaccharide is attached to the bacterial cell surface.

Although not impossible, the ability to produce and isolate new polysaccharides with modified structures from bacteria by conventional microbiological methods requires the following tedious approach. Strains producing potentially useful materials are isolated from the wild and converted to high yield strains by mutagenesis. This has been accomplished for capsular polysaccharides in a wide range of bacterial species[5,6,7]. Isolation of new polysaccharide structures requires extremely large scale screening of bacterial populations to isolate rare mutants with structurally-modified polysaccharides. The rarity of structural variants is due to the fact that polysaccharides are the end-product of complex multi-step enzymatic biosynthetic pathways. As such they are not readily amenable to modification by recombinant DNA technology. Typically, mutations, which result in the production of new structures, are too rare for the conventional screening approach to the problem to be viable.

Our laboratory has developed (and is in the process of refining) new and generally applicable techniques to produce and isolate high yields of structurally-modified polysaccharides and polysaccharases. We have utilized the interaction between bacteriophage and bacteria for the purification, processing and genetic modification of bacterial polysaccharides and their basic oligosaccharide repeat units. Bacteriophage, or simply phage, are virus like particles that infect and kill bacteria. The bacteriophage we have chosen to study distinguish themselves by producing a polysaccharide degrading enzyme, endoglycanase, which is specific for the polymer produced by the bacterial strain they infect. As a result, the bacterial polysaccharide capsule that would otherwise act as a physical barrier to phage infection is disrupted, permitting infection and resulting in cell death. Understanding the details of this process has lead to the development of a general set of biochemical and genetic techniques for the manipulation of polysaccharides. These techniques can be used to process and produce naturally occurring materials with desirable physical, mechanical, and biological properties.

APPROACH

The basic bacteriophage-based strategy for adapting bacteria for polysaccharide production is as follows. A strain producing a naturally occurring material with the property of interest is identified. If necessary, established mutagenesis techniques are used to produce a high yield polymer producing strain. Bacteriophage enzymes can be used to detach polymer product from the cells which produce them. Cells are then readily removed by a short low speed centrifugation leaving the polymer produced in solution. If desired, the material can be further processed to a high molecular weight form (greater than 500,000) by partial digestion of the material with purified bacteriophage enzyme. This yields material suitable for thin film fabrication. Alternatively, the material may be digested to its basic repeat oligomeric structure, with yields in excess of 50% and of high stereospecific purity due to the nature of the synthesis. If one seeks to alter the property of the native polymer, a population of bacterial cells is exposed to bacteriophage bearing endopolysaccharase specific to the native capsular polysaccharide. Those bacteria which are resistant to the phage often survive by virtue of producing a structurally modified capsular polysaccharide with reduced susceptibility to the bacteriophage enzyme.

The process we have developed to produce and isolate bacterial-derived polysaccharides can be illustrated by our search for polymers with "drag-reducing" properties.

The phenomenon of drag-reduction by suppression of turbulent flow in pipes has been studied for over forty years. Recently, this poorly understood non-newtonian phenomenon has been linked to the extensional viscosity of the active polymer[8]. These polymers have the ability to suppress turbulent flow at Reynolds numbers above the

transition point[9]. Several researchers have identified polysaccharides as polymers that have drag-reducing properties[10,11]. In fact, xanthan gum, a common industrial polysaccharide of bacterial origin is recognized as among one of the most effective drag-reducing polymers on a weight/ppm basis.

To test our process of producing and isolating bacterially-derived polysaccharides we have obtained a copy of the World Health Organization's *Klebsiella pneumoniae* serotypes, which include 79 closely related strains that produce capsular polysaccharides of known chemical structure. These strains were screened for production of drag-reducing polymers.

Several *K. pneumoniae* strains were found to produce drag-reducing material strain K63 produced one of the most effective polysaccharides. The native K63 polysaccharide, with its basic repeat unit of [galactose-acetylated galacturonic acid-fucose]$_n$, was found to be approximately twice as effective as a drag-reducer as xanthan gum[12]. The polymer also showed a high degree of resistance to mechanical shearing, as measured by repeated passage through a turbulent flow rheometer. Without bacteriophage-enzyme treatment it was impossible to effectively separate the polysaccharide from cells, even at centrifugation speeds of 50,000 x g for 4 hours. Brief bacteriophage-enzyme treatment, however, permitted rapid cell polymer separation of high molecular weight material (approximately 3,500,000), after 20 minutes of centrifugation at 6,000 x g. Furthermore, the ability to remove polysaccharide from the cell surface lead to an eighty percent increase in polymer recovery.

As depicted in Figure 1, further bacteriophage-enzymatic processing of the purified material, under precisely controlled conditions, leads to a reproducible partial digestion of this material. This resulted in the production of a homogeneous mono-disperse polysaccharide of average 900,000 molecular weight. This material was suitable for uniform thin-film formation (25 nanometers). Alternatively, complete limit digest of the material lead to high yield (64%) conversion of the native polymer to its basic repeat trimer. In this manner, gram quantities of oligosaccharide, with purity sufficient for crystallization, were produced.

Enzyme-bearing bacteriophage, which specifically infects *K. pneumoniae* K63 has been isolated and produced in large quantity. The bacteriophage's ability to specifically degrade K63 capsular polysaccharide has been used as a selection tool to generate a family of polysaccharides with related structures. Among the bacteria surviving this exposure to this bacteriophage are isolates which could be readily identified as having altered polysaccharide structure (see Table 1). These mutants showed a reduced level of non-stoichiometric pyruvylation of the capsular polysaccharide. Nuclear magnetic resonance (NMR) studies clearly identified mutants with changes in this structure of the polysaccharides repeat trimer.

As expected, the mutants with structurally altered polysaccharides showed altered rheological properties. In fact there was a significant improvement in both their drag-reducing effectiveness (over 10%) and mechanical, or shear stability (also above 10%). Furthermore, this material proved to be enzymatically convertible to oligosaccharides so that a new family of closely related structural oligomers has also been generated.

CONCLUSION

Bacterial polysaccharides represent a diverse and largely untapped source of polymeric materials and specialty chemicals, including possible starting materials for pharmaceuticals. The general area of adapting polysaccharides produced from bacterial sources is hampered economically by the high cost of purifying polysaccharides for commodity chemical uses and the limited ability to structurally modify native polymers to enhance performance[1].

Enzyme-bearing bacteriophages with endopolysaccharase activity have been demonstrated to be useful tools for polymer purification and processing. Furthermore, bacteriophage can be used as a selective agent to generate families of polysaccharides structurally related to native material, but with altered properties. In this way, new polysaccharides with enhanced properties can be generated with high probability. This report demonstrates the application of enzyme-bearing bacteriophage to select mutants of *K. pneumoniae* K63 with enhanced rheological performance.

We have identified the production of oligosaccharides for pharmaceutical biosynthesis as a possible important spin-off of our ongoing material science research. It should be clear that the bacteriophage enzyme techniques outlined here permit the generation of, not only new and novel, polysaccharides, but also large quantities of new oligosaccharides. Gram quantity production of these compounds in our laboratory has become routine.

Generation of oligosaccharides by the alternate route of bacteriophage enzymatic degradation of bacterial polymers, coupled with the ability to produce families of polysaccharide structural variants, provides a promising new source of materials for synthesis of biologically active carbohydrates.

ACKNOWLEDGEMENTS

The work described in this report was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The work was supported under contract with the Defense Advanced Projects Agency, Department of Defense.

REFERENCES

1. Aspinal, G.O. The Polysaccharides. Harcourt, Brace, and Jovanovich, 1982.
2. Yalpani, M. Industrial Polysaccharides: Genetic Engineering Structure/Property Relations and Applications. Elsevier, 1987.
3. Toone E.J. and Whitesides, G.M. in Enzymes in Carbohydrate Synthesis, Bednarski M.D. and Simon E.S. eds. ACS, 1991.
4. Weiss R.M. and Ollis, D.F. Biotechnology and Bioengineering: 22:859-873, 1980.
5. Sutherland, I.W. J. Applied Biochem: 1:60-70, 1979.
6. Lim, S.T., Uratsu, S.L. and Kern, R.G. Microbios Letters: 29:121-125, 1985.
7. Cadmus, M.C. Rogovin, S.P. Burton, K.A. et al. Can J. Microbiol: 22:942-948, 1976.
8. Petersen, G.R., Nelson, G.A., Cathey, C.A., and Fuller, G.G. Applied Biochemistry and Biotechnology: 20/21:845-867, 1989.
9. Virk, P.S. AIChE Journal: 21:625-656, 1975.
10. Hoyt, J.W. Trends Biotechnology: 3:17-21, 1985.
11. Petersen G.R. Schubert, W.W., Richards, G.F. and Nelson, G.A. Enzyme Microb. Technol.:12:255-259, 1990.
12. Kern, R.G. AIP Conference Proceedings: 137:135-142, 1985.

Enzymatic Processing of Polysaccharides Yields Material Suitable for Further Processing / Manipulation

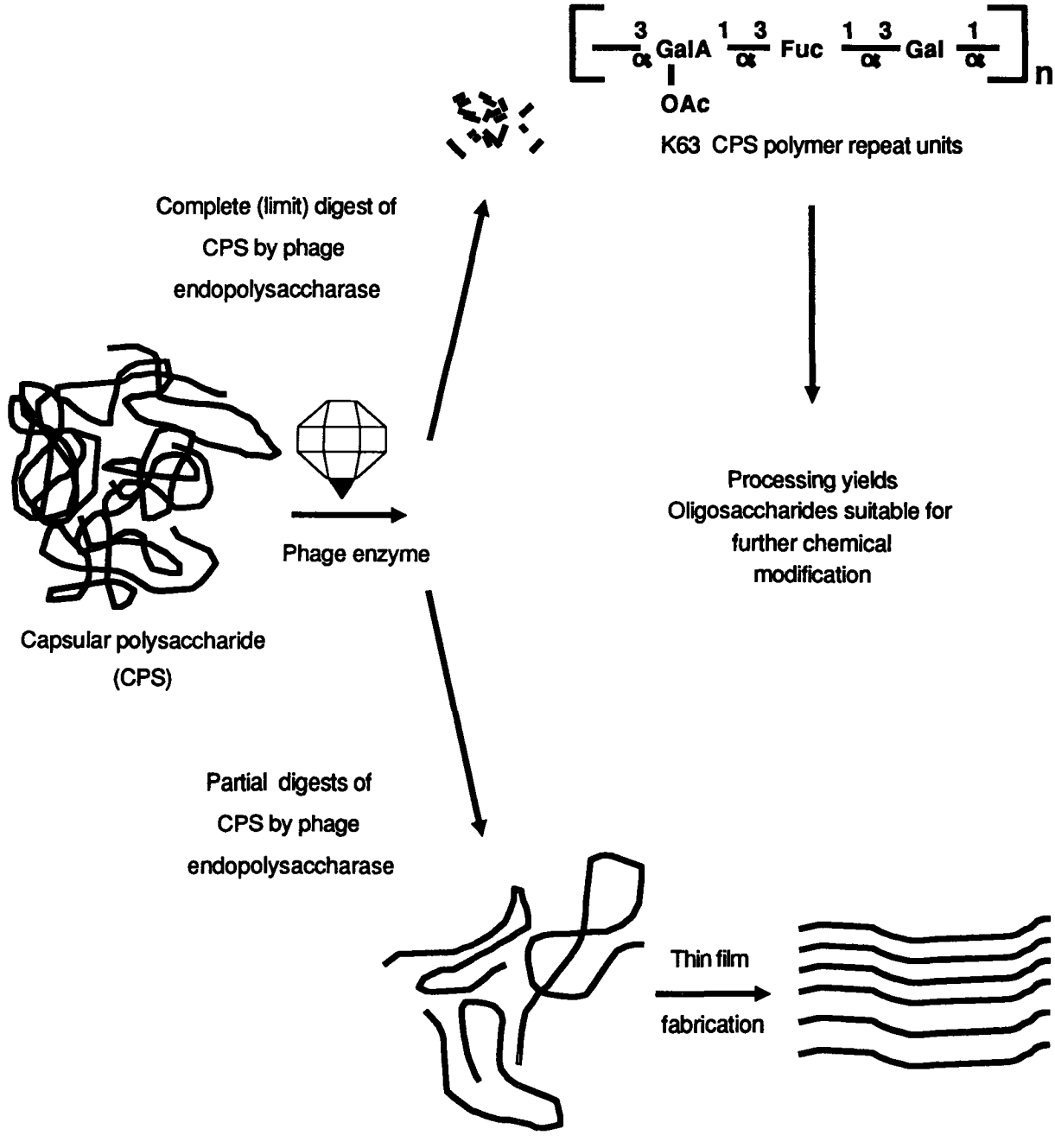


FIGURE 1

Table 1 Summary of Properties of *K. pneumoniae* K63 Variants

STRAIN	% PYRUVALATION[1]	NMR DONE = + Δ = difference	% DRAG REDUCTION[2]	% SHEAR STABILITY[3]
K63 wt	0.83	+	55.9	84.9
K63-JPL1	0.11	+ Δ (?)	67.7	95.6
K63-JPL2	0.04	+	66.7	95.6
K63-JPL3	0.11	+	67.0	99.4
K63-JPL4	0.12	+	67.6	98.2
K63-JPL5	0.11	+ Δ (GalA)	67.0	99.4
K63-JPL6	0.064	+	67.6	95.7
K63-JPL8	0.19	+	67.2	99.7
K63-JPL9	--	+ Δ (Fuc)	--	--

1. As % dry weight
2. % Drag-reduction at 25 wppm
3. % Drag-reduction of 25 wppm solution retained after three passes through rheometer

N92-22730

CRYOGENIC FOCUSSING, OHMICALLY HEATED ON-COLUMN TRAP

Stephen R. Springston
Environmental Chemistry Division
Department of Applied Science
Brookhaven National Laboratory
Upton, NY 11973

ABSTRACT

A procedure is described for depositing a conductive layer of gold on the exterior of a fused-silica capillary used in gas chromatography. By subjecting a section of the column near the inlet to a thermal cycle of cryogenic cooling and ohmic heating, volatile samples are concentrated and subsequently injected. The performance of this trap as a chromatographic injector is demonstrated. Several additional applications are suggested and the unique properties of this device are discussed.

INTRODUCTION

The importance of temperature and its effects on physico-chemical processes cannot be overemphasized. With few exceptions, process rates increase with the temperature. In gas chromatography (GC), a linear increase in temperature exponentially raises the vapor pressure of solutes partitioned between the stationary and mobile phases. The sequential elution of progressively less volatile, and generally larger, solutes by steadily increasing the column temperature has long been practiced in GC. Thus, samples with a broad range of individual solute volatilities can be analyzed in a single separation. Conversely, dilute vapor samples may be concentrated as a narrow band by passing the sample stream through a cooled region. This trapping region is quickly warmed to release the material in a sharply defined pulse where the solute concentration has been greatly enhanced over the original dilute stream. A novel solute trap for capillary chromatography has been recently developed in this laboratory which controls the temperature in a short section of the analytical column. By carrying out the concentration within the column itself, many problems associated with precolumn concentration schemes are avoided. The design, construction, evaluation, and use of a chromatographic trap based on an ohmically-heated column section illustrates the unique characteristics of this technology. Many of these same features also suggest other novel applications, including some outside the field of chromatography.

In elution chromatography, sample mixtures must be introduced into the analytical column as a zone which is spatially narrow relative to the separated zones of individual components which later emerge. A second requirement dictates that the sample quantity introduced must be detectable. Numerous injection devices have been developed which meet these criteria (1). An appropriate technique should be selected based on the sample and its accompanying matrix. Cryogenic trapping is one strategy particularly well suited for samples which must be concentrated prior to separation by GC.

The principle of cryogenic trapping is straightforward. Chilling a region of the flow stream lowers the sample vapor pressure and causes solutes to be concentrated in a stationary zone. Following the concentration step, the trapped material is then heated to quantitatively transfer the sample into the separating column. It is important to heat the trap quickly so that all material leaves as a narrow band. Under proper conditions, sample concentrations in this discrete plug may be increased four or more orders of magnitude above their original level. Two points are crucial in the design of a cryogenic trap. First, the trap volume must be small, approximately the same size as the ultimate injection volume. Second, the trap must be capable of rapid heating to quickly vaporize the trapped material and avoid drawing out the narrow injection zone.

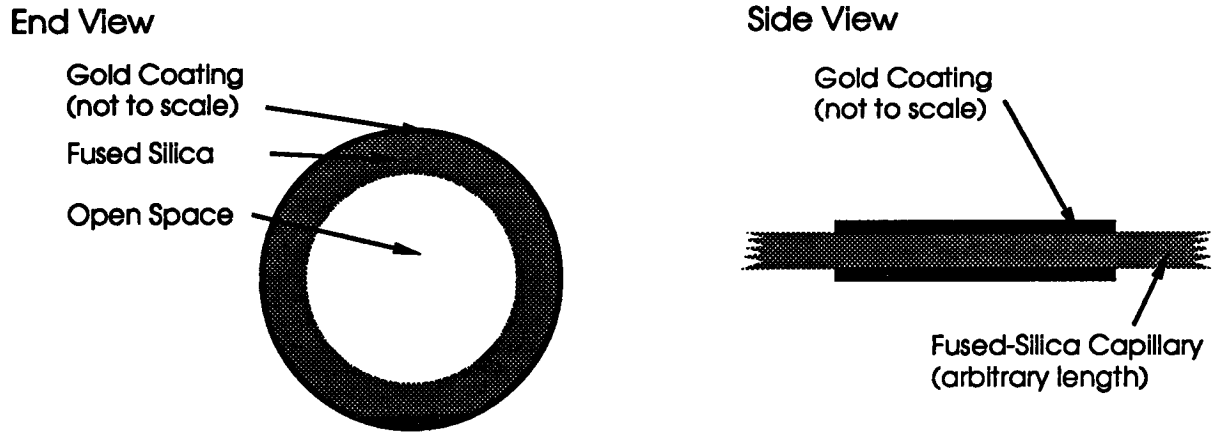


Figure 1. Schematic of a gold layer on a capillary column. For clarity, the polyimide layer and stationary phase are not shown.

PROCEDURE

Fabrication of the on-column trap entails depositing a thin layer of gold on the outside of a fused-silica capillary column about 10 cm from the inlet and extending for a length of 10 cm. Any commercial capillary column is suitable for operation with this trap. This evaluation was done with 0.25-mm i.d. columns which were statically coated (on the inside) with a 0.25-μm thick coating of SE-30 (dimethyl silicone) and showed a coating efficiency of 93% for dodecane at 100°C. Two lengths, 11.0 and 1.92 m, were used to isolate any diminishment of column efficiency due to trapping. The gold layer is applied as a gold complex dissolved in volatile oils (Liquid Bright Gold, No. 7621, Engelhard, Hanovia Liquid Gold Division, East Newark, NJ). The solution is applied with a brush and dries to an even coating in air. Once dry, the entire column is placed in an oven and heated to 310°C for 20 min. Heating drives off the residual solvent and decomposes the organic portion of the gold complex leaving a specular layer of metallic gold on the outer capillary wall. The application and firing process is repeated three times to yield a gold layer about 0.3 μm thick. During heating, the column interior is purged with helium to remove any trace decomposition products which may evolve from the stationary phase. The outer polyimide coating of the capillary column is unaffected by the gold layer. Because this gold layer is so thin, the remarkable flexibility of fused-silica columns is retained.

Figure 1 shows a schematic representation of the trap on a capillary column. A polyimide layer is applied to the outside of fused-silica tubing during the manufacturing process to protect against damage. The trap is easily visible over the polyimide layer and resists abrasion. When not in use, the trap does not alter the performance of the column in any way. Ferrules seal equally well to gold and polyimide, however, ferrules are easier to remove from the gold layer because they stick less.

Capillary columns outfitted with a gold trap were installed in a gas chromatograph (Varian VISTA 4600, Varian Instruments, Walnut Creek, CA) configured as shown in Figure 2 for purposes of evaluation. A split/splitless injector allowed the trap performance to be characterized and compared with conventional injection techniques. For routine analyses with the cryogenic trap, a sample loop is substituted for the split/splitless injector. Electrical connections to the gold layer are made by wrapping conductive fibers around both ends of the trap. A variable voltage source regulates the potential across the trap. The voltage and current are monitored simultaneously. The trap is bent such that the region between the electrical leads is held above liquid nitrogen contained in a Dewar. The trap and its lead connections remain electrically isolated from the Dewar and gas chromatograph. The trapping mode is actuated by turning off the voltage source and allowing the trap to cool to -150°C. When a voltage is applied, the trap temperature quickly rises and samples are driven off the trap. For the traps described here, voltages between 10 and 15 V were used. Due to the thinness of the gold layer, the trap resistance is relatively high and the current is on the order of 100 mA.

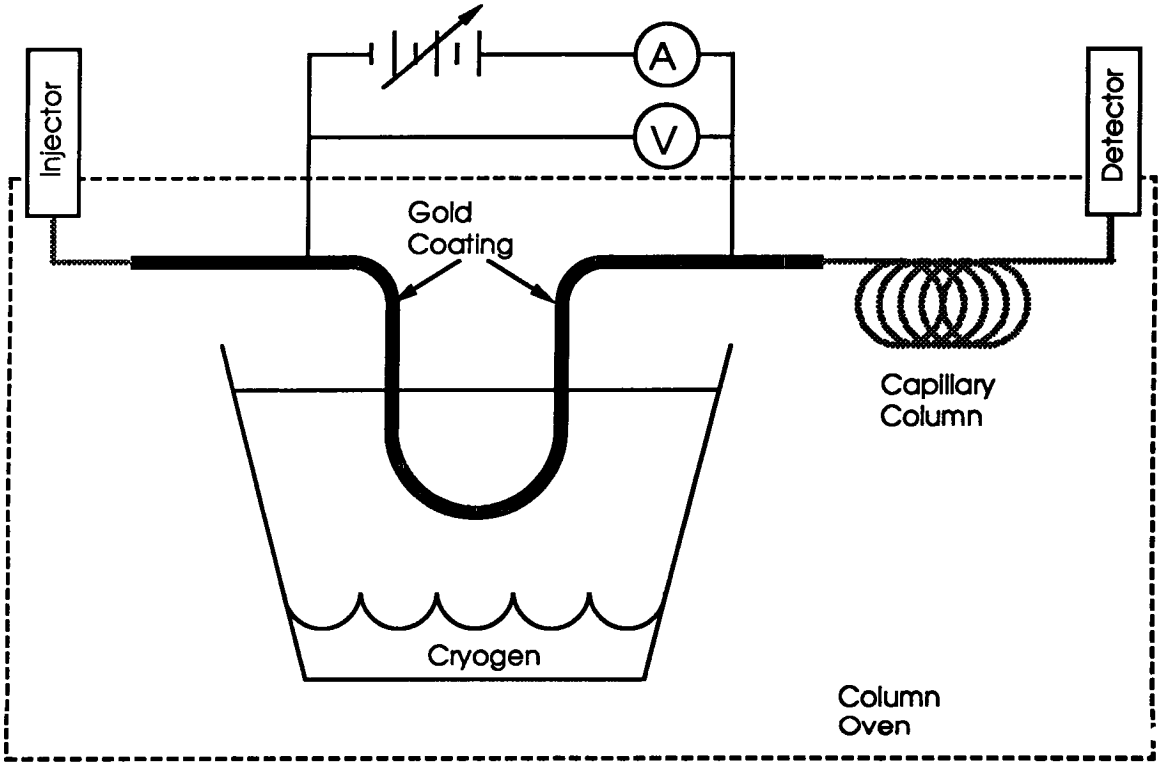


Figure 2. Schematic diagram of capillary trap installed in gas chromatograph.

Separated solutes were detected by a flame ionization detector (FID) as they elute from the column. For the purposes of this study, moment analysis (2) was used to accurately characterize the shape of solute peaks. The detector output was stored digitally at 20 Hz.

EVALUATION

Operation of the gold-coated capillary section as a trap for capillary chromatography is illustrated by three demonstrations. First, the temperature of the trap can be accurately monitored during operation by simply measuring the electrical resistance. Most materials exhibit increased electrical resistance as the temperature rises. The resistivity of pure gold has a temperature coefficient of 0.0034 per °C (3). This relationship was confirmed experimentally for these traps by directly measuring the internal temperature using a miniature iron-constantan thermocouple (Omega Engineering, Stamford, CT) carefully inserted into a capillary. While the GC oven temperature was raised, the trap resistance was measured. Next, with the oven off, the trap was ohmically heated by increasing the current in a step-wise fashion. By also measuring the voltage across the trap during ohmic heating, the trap resistance was calculated using Ohm's law and recorded as a function of the directly measured temperature.

The line in Figure 3 shows the resistance across the trap as a function of temperature while the chromatograph oven was used to heat the capillary. The step-wise nature of the line is due to digitization error in measuring the resistance. The open circles correspond to the trap resistance calculated from Ohm's law as the trap is ohmically heated. In both cases, the trap resistance increases linearly with temperature. The apparent temperature coefficient of resistance is 0.001 per °C and is independent of whether the trap is heated externally or ohmically. Resistance in the leads and connections accounts for the difference between the experimental and reference values. Note the tedious operation of inserting the thermocouple is performed here

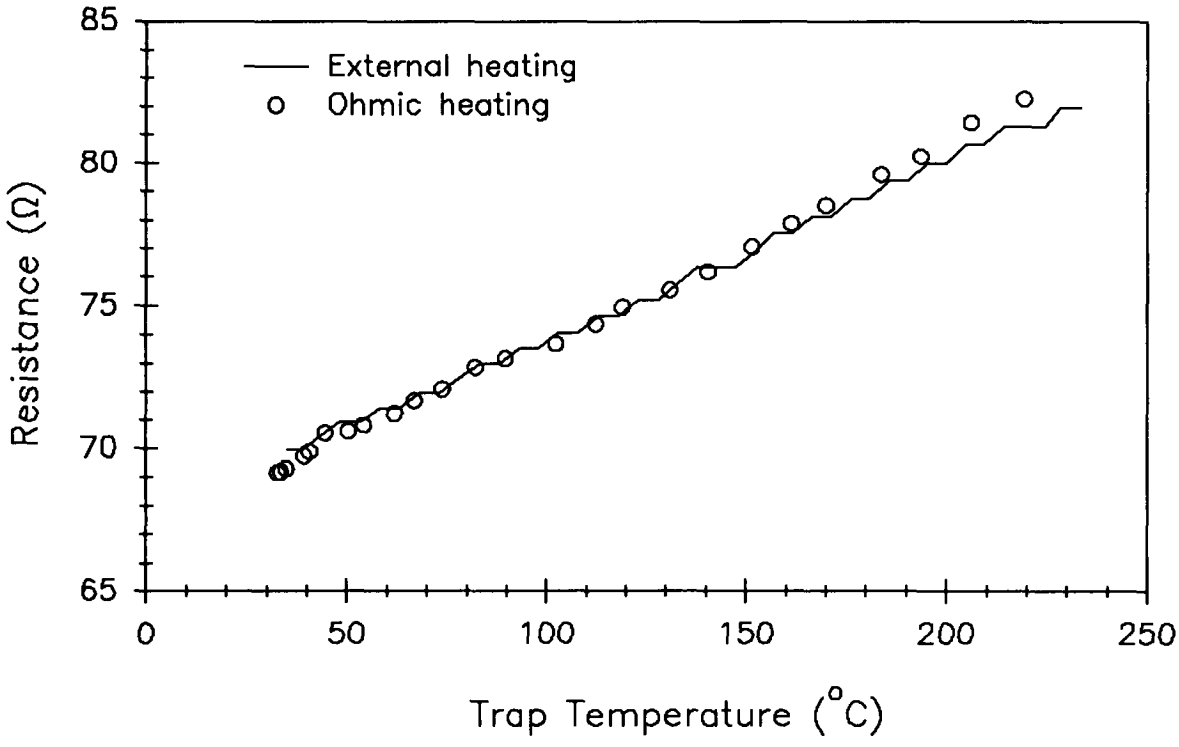


Figure 3. Trap resistance *vs.* trap temperature. The trap temperature was measured inside the capillary.

only to establish that the effective coefficient of resistance can be measured by external heating. In actual use, a new trap is calibrated merely by programming the oven temperature while measuring the resistance.

Second, the ability of the trap to quantitatively capture and release volatile samples is demonstrated using the 11-m capillary. This length was chosen to provide complete separation of the C5 - C8 normal alkanes. A stock solution of 2000 parts-per-million-volume (ppmv) in room air was sequentially diluted to 50-, 100-, and 200-parts-per-billion-volume (ppbv) in air. Figure 4A shows a separation of the stock solution using a conventional split injector. The small injection volume and the 99% of sample dumped to the vent define the amount of sample actually reaching the column and detector to be 300 pg/component. These losses are necessary in split injections to provide a narrow injection band. Figure 4B shows a separation of the stock solution diluted by a factor of 40,000 (50 ppbv) injected with a trap. The sample volume has been increased to 1 mL. In this case, there is roughly 150 pg/component on column. The additional peaks in Figure 4B are due to various compounds in the air used to dilute the stock solution. Methane, though present in ambient air at 20 times the level of the other alkanes is not concentrated at the trap temperature and, therefore, is not detected.

The trapping period was two minutes in this case, but longer periods (and larger sample volumes) can be accommodated easily. The chromatogram in Figure 4B indicates where heating commenced, thus beginning the separation. This corresponds to the injection time for a split injection. The uniformity of retention times is evident from Figure 4. Statistical evaluation of repeated injections has shown retention times for trapped samples are reproducible to <0.3 s and indistinguishable from split injections (4).

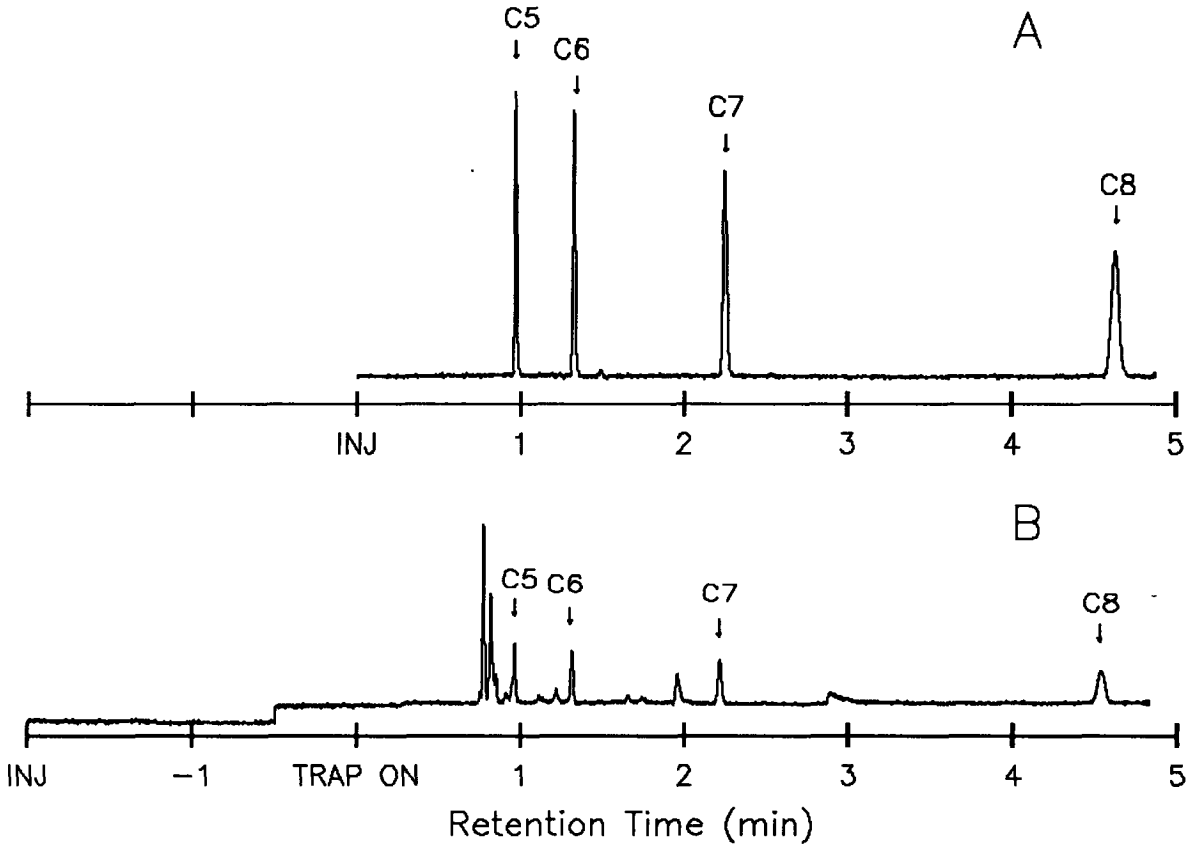


Figure 4. Chromatograms of pentane, hexane, heptane, and octane on an 11 m x 250-μm i.d. capillary column coated with 0.4 w/v% SE-30. (A) 5-μL vapor-phase injection of 2000 ppmv, split 100:1. (B) 1-mL vapor phase injection of 50 ppbv, no split, trapped for 2 min.

Quantitative trapping is demonstrated in Figure 5 as a plot of peak area *vs.* concentration for two solutes of different retention ratios. Error bars indicate $\pm$ 1 standard deviation. Digitization error for these small signals contributes to the scatter. The precision for all solutes at 50 ppbv is better than 10% RSD. Replicate split injections with larger signals show a precision of 3% RSD. As with any injection technique, referencing peak areas to an internal standard would greatly enhance the precision of quantitation.

Third, the effects of trapping on peak shapes are illustrated. As stated earlier, samples must be delivered to the analytical column in a narrow concentration pulse. To measure the ability of the trap to release samples as a well-defined pulse, a 1.92-m column was used. This short column length maximizes any contributions to peak variance from the injection technique. A slow release during the heating step would be readily apparent as broad peaks in the resulting chromatogram. For comparison, a 100:1 split injection provides an "ideal" sample inlet system.

The parameters used to characterize the peak shapes were the second and third statistical moments calculated from the digitized chromatograms. The second moment (or variance) is related to the peak width and increases with retention time for isothermal separations. Significant departure from an ideal injection profile

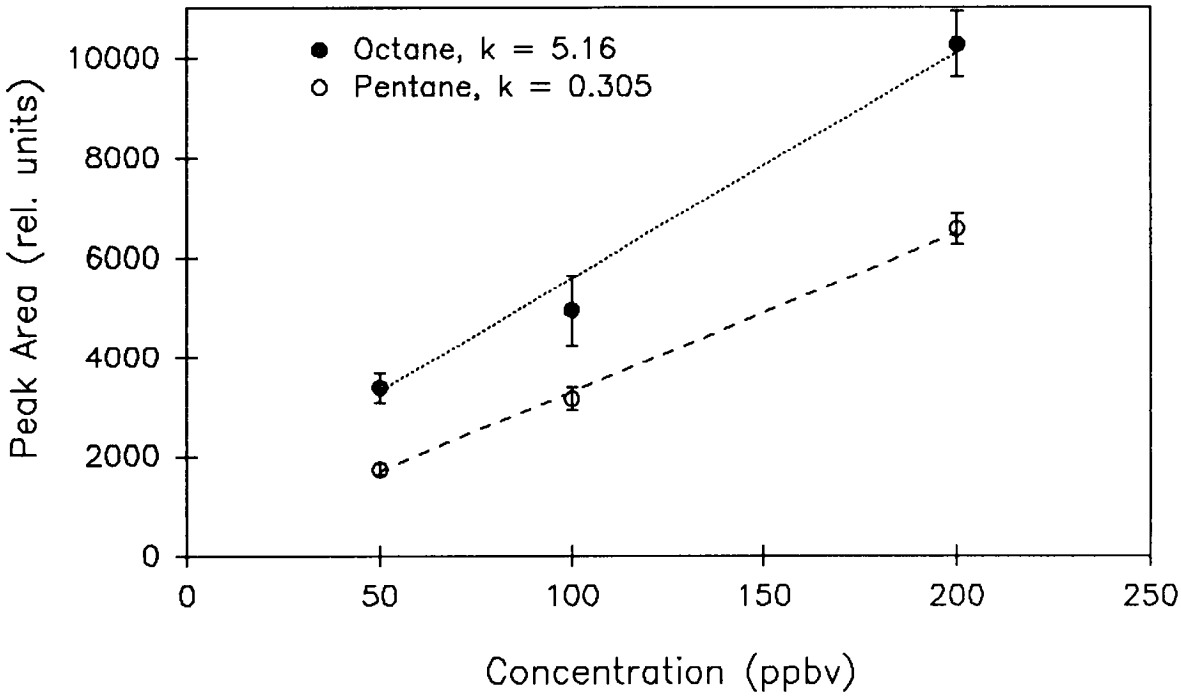


Figure 5. Peak area *vs.* concentration for solutes on the column of Fig. 4. Samples were trapped for 2 min. Mean ± 1 standard deviation.

causes the second moment to increase relative to an ideal injection. The third moment is a measure of peak asymmetry and is zero for an ideal peak.

The behavior of the column trap relative to a split injection is shown in Table I. The second and third moments for peaks from a trapped injection compare favorably with the "ideally" shaped peaks from a split injection. (Heptane exhibited a very minor impurity which degraded its peak shape.) Because longer columns are less demanding of the injection profile, these results, obtained under artificially demanding conditions, confirm the suitability of this trap for chromatographic purposes.

TABLE I

Second and third statistical moments for split and trapped injections on a 1.92-m column.

	2nd Moment x 10^2 (sec^2)		3rd Moment x 10^2 (sec^3)	
Sample	Split	Trapped	Split	Trapped
C5	6.11 ± 0.49	5.33 ± 0.19	2.31 ± 0.48	1.78 ± 0.15
C6	8.68 ± 0.59	7.81 ± 0.55	4.26 ± 0.64	3.58 ± 0.71
C7	9.39 ± 1.81	9.95 ± 0.86	3.64 ± 2.96	5.30 ± 1.14
C8	20.59 ± 0.55	17.56 ± 0.73	2.68 ± 0.34	2.40 ± 0.29

Mean ± 1 standard deviation (n = 7).

POTENTIAL APPLICATIONS

In the future, the temperature in the trap may be accurately controlled using a simple feedback circuit. Such a circuit would allow trapped compounds to be sequentially removed by programmed heating. Provided sufficient differences in volatility exist, unwanted compounds could be back-flushed from the trap and prevented from entering the column and detector. For example, water present in atmospheric samples could be eliminated from the chromatographic analysis. This would greatly speed the analytical throughput.

The device described in this report shows potential far beyond its use as an injection device for gas-phase elution chromatography. Due to its low thermal mass and high thermal slew rates, the trap is ideal for multiplex chromatography (5), an approach where samples are repeatedly injected at intervals less than the analysis time of an individual injection. The resultant detector output is then mathematically deconvoluted to provide continuous (rather than batch) concentration information.

In supercritical fluid chromatography (SFC) (6), the mobile-phase density controls solute retention. A capillary trap could be used to increase the temperature in a localized column section, thus trapping non-volatile solutes on the walls. To inject the sample, the current would then be turned off and the trap allowed to cool. This reverses the protocol for injecting trapped samples in capillary GC as described above. Using a thermal trap for SFC was suggested several years ago (7), and preliminary results using other heaters have been reported (8). Use of the ohmically-heated trap described here is being evaluated.

Another potential application related to SFC is the construction of a controllable flow restrictor. Drawn out capillary tips are used to maintain back pressure in columns upstream of atmospheric pressure detectors (9). These tips generally offer a fixed restriction which differs from that required for optimal flow. Furthermore, density programming is often used to enhance separations, causing the flow rate to change during an analysis and making it impossible to maintain the ideal mobile-phase velocity. Adjusting the tip temperature would alter the location where the supercritical-fluid mobile phase converts to a gas. Because gases have about the same viscosity as fluids, but are two orders of magnitude less dense, warming the tip increases the back pressure generated in a trap. Regulation of the restrictor temperature has been advocated as a method to control column flow by Berger and Toney (10). From theoretical considerations, Berger concludes that high temperatures and short heated zones are required for SFC (11). This author goes on to declare that poor thermal contact in ordinary heaters make heat transfer into a fluid difficult. The possibilities of using an ohmically-heated, gold-coated restrictor are intriguing in view of the excellent heat transfer between the gold layer and the internal passage. The integral thermometric feature of the gold layer should permit accurate control of the flow through the analytical capillary, even during pressure programming.

A section of gold-coated capillary could function as an inert, fast response, flow controller for gases. As the temperature increases from 0° to 300°C, most gases vary in viscosity by a factor of two. A short section of capillary inserted in an otherwise unrestricted flow stream acts as a limiting restriction. Changing the temperature in this restriction by ohmic heating would then change the flow for the entire system. This technique would be practical only for controlling flows over a modest range, but the fast response, lack of moving parts, and the inert chemical nature of the wetted surfaces are attributes which would distinguish such a controller from its mechanical counterparts.

SUMMARY OF ADVANTAGES

The performance of the ohmically-heated trap as an injector for capillary GC has been demonstrated in a preliminary fashion. This trap exhibits a unique combination of features which should prove useful for other applications, some of which have been proposed in this paper. A review of these features may suggest additional applications.

Minimal Thermal Mass - The amount of gold applied is very small and contributes little to the thermal mass over the length of the trap. Because the total combined thermal mass of capillary and gold is low, rapid

thermal slew rates are possible with low power input as shown in the experiments above. Rapid cooling is also facilitated. Consumption of cryogen is minimal in cycling between heating and cooling.

No Moving Parts - In the configuration used for this study, the trap remained exposed to cryogen, even during ohmic heating. No valves were needed to switch flow streams. The current was remotely turned on and off with a relay.

No Compromise with Other Injection Methods - The gold layer does not preclude the use of any other injection technique if on-column trapping is not desired. In this case, the column containing the trap is merely connected to the injector as an ordinary capillary column.

Connection Free - The trap is formed as an integral part of the analytical column. Unlike other purge and trap systems, there are no mechanical junctions to broaden the injection pulse as it moves from the trap into the separation portion of the column. No on-column focussing, such as temperature programming, is necessary. Isothermal separations are particularly useful for highly volatile samples. In high-pressure systems, reducing the number of connections diminishes the possibility of leaks.

Inert Wetted Surface - Because the trap is deposited on the column exterior, samples only contact the inert, interior surface of the capillary. The elimination of fittings encountered by the concentrated sample also minimizes the opportunity for catalytic decomposition of injected components. This feature is crucial for labile compounds.

Applicable to Existing Columns - Virtually any capillary column may be fitted with a capillary trap. The application of the gold layer is relatively simple and may be done with readily available equipment. In cases where the stationary phase would not withstand the temperature needed to fire the gold layer, the column should be purged with an inert gas from the detector end while heating only the trap section.

Efficient Heating - Unlike metal tubes or coils of resistance wire, the gold layer is so thin that the electrical resistance is on the order of 100 Ω. Due to the low thermal mass and intimate contact between the gold and capillary, only 1 - 2 W of power is needed to heat a trap to 300°C, even when the surrounding gas is at -150°C. The currents required (50 - 150 mA) are quite modest and can be carried by fine connecting wire (which maintains the low thermal mass).

Homogeneous Temperature - The liquid gold solution is designed to cover surfaces uniformly. Properly applied, the gold layer is also highly uniform. Within the precision of the thermocouple used (± 1°C), no variation in temperature was observed over a 10-cm length during heating.

Integral Thermistor - As shown above, the resistance across the trap is linearly related to the trap temperature over a wide range. An external temperature sensor, which would add to the thermal mass, is not required.

Wide Temperature Range - The gold layer does not, by itself, impose any temperature limits on the trap. The minimum trapping temperature for this trap is determined by the cryogen. Liquid nitrogen is satisfactory for virtually all organic compounds. Above 350°C, the polyimide on fused silica begins to oxidize and the column becomes extremely brittle.

Controllable Temperature - Any temperature over the available range can be maintained with a constant voltage source. Temperature control would be greatly enhanced by using a feedback system to adjust the voltage and achieve a constant resistance. The low power requirements of this trap are easily supplied without elaborate circuitry.

Low Cost - The material needed to prepare a trap costs less than $1.00. (Most of this is wasted by inefficient application.) Coating the entire capillary length would not significantly increase the cost of an

analytical column and could provide a high-temperature alternative to the polyimide layer currently protecting most fused-silica capillaries.

ACKNOWLEDGEMENT

This research was performed under the auspices of the U.S. Department of Energy under Contract No. DE-AC02-76CH000016.

REFERENCES

1. J.V. Hinshaw, *J. Chromatogr. Sci.*, 26 (1988) 142.

2. S.N. Chesler and S.P. Cram, *Anal. Chem.*, 43 (1971) 1922.

3. R.C. Weast (ed.), "Handbook of Chemistry and Physics", The Chemical Rubber Co., Ohio (1972) 53rd Edition, p. E72.

4. S.R. Springston, *J. Chromatogr.*, 517 (1990) 67.

5. J.B. Phillips, D. Luu, J.B. Pawliszyn and G.C. Carle, *Anal. Chem.*, 57 (1985) 2779.

6. R.D. Smith, B.W. Wright and C.R. Yonker, *Anal. Chem.*, 60 (1988) 1323A.

7. S.R. Springston, *Ph.D. Thesis*, Indiana University, Bloomington, IN, 1984, p. 46.

8. M.L. Lee, B. Xu, E.C. Huang, N.M. Djordjevic, H-C.K. Chang and K.E. Markides, *J. Microcolumn Sep.*, 1 (1989) 7.

9. S. Green and W. Bertsch, *J. High Resolut. Chromatogr. Chromatogr. Comm.*, 11 (1988) 414.

10. T.A. Berger and C. Toney, *J. Chromatogr.*, 465 (1989) 157.

11. T.A. Berger, *Anal. Chem.*, 61 (1989) 356.

N92-22731

XPS STUDY OF THE EFFECT OF HYDROCARBON CONTAMINATION ON POLYTETRAFLUOROETHYLENE (TEFLON) EXPOSED TO ATOMIC OXYGEN

Morton A. Golub and Theodore Wydeven
NASA Ames Research Center
Moffett Field, California 94035

and

Robert D. Cormia
Surface Science Laboratories
1206 Charleston Road
Mountain View, California 94043

ABSTRACT

The presence of hydrocarbon contamination on the surface of polytetrafluoroethylene (PTFE) markedly affects the oxygen uptake, and hence the wettability, of this polymer when exposed to an oxygen plasma. As revealed by XPS (X-ray photoelectron spectroscopy) analysis, the oxygen-to-carbon ratio (O/C) for such a polymer can increase sharply, and correspondingly the fluorine-to-carbon ratio (F/C) can decrease sharply, at very short exposure times; at longer times, however, such changes in the O/C and F/C ratios reverse direction, and these ratios then assume values similar to those of the unexposed PTFE. The greater the extent of hydrocarbon contamination in the PTFE, the larger are the amplitudes of the "spikes" in the O/C- and F/C-exposure time plots. In contrast, a pristine PTFE experiences a very small, *monotonic* increase of surface oxidation or O/C ratio with time of exposure to oxygen atoms, while the F/C ratio is virtually unchanged from that of the unexposed polymer (2.0). Unless the presence of adventitious hydrocarbon is taken into account, anomalous surface properties relating to polymer adhesion may be improperly ascribed to PTFE exposed to an oxygen plasma.

INTRODUCTION

Morra and co-workers (1) recently reported that the surface of the important polymer Teflon, or polytetrafluoroethylene (PTFE), when exposed to an oxygen plasma for very short times exhibited a sharp increase in the oxygen-to-carbon ratio (O/C, increasing from 0.014 to 0.129), and a correspondingly sharp decrease in the fluorine-to-carbon ratio (F/C, decreasing from 1.73 to 1.26), as measured by electron spectroscopy for chemical analysis (ESCA) (or X-ray photoelectron spectroscopy (XPS)). At longer exposure times, however, the changes in the O/C and F/C ratios reversed direction, and these ratios assumed values similar to those of the unexposed PTFE. These ESCA results proved surprising to us since they conflicted with our prior observation (2) that PTFE experienced "a very small [but monotonic] increase of surface oxidation with time of exposure to $O(^3P)$ in an r.f. O_2 discharge [while] the F/C ratios were virtually unchanged from that of the control (2.0)." From a close examination of the ESCA spectra and data presented by Morra and co-workers, we suspected that the spikes observed in their plots of O/C or F/C versus time of exposure to an O_2 plasma were not characteristic of PTFE per se but were instead a result of the hydrocarbon contamination present in their PTFE samples. To be sure, those workers noted that the ESCA spectra of their PTFE, both before and after exposure to an O_2 plasma for 15 min, showed "a weak structure due to hydrocarbon contamination," but they tacitly assumed that this hydrocarbon played no role in the effect of exposure time on the O_2 plasma-induced surface modification of PTFE. We subsequently demonstrated that, on the contrary, the spikes observed in the O/C- and F/C-exposure time plots presented by Morra and co-workers were a direct result of the hydrocarbon contamination present in their PTFE sample (3). Indeed, we pointed out that this contamination was not minor, nor could their untreated polymer be considered to have a *clean* PTFE surface, inasmuch as its F/C ratio (1.73) was well below the theoretical value of 2.0.

DISCUSSION

Figures 1 and 2 present composite plots of O/C and F/C ratios, respectively, as a function of time of exposure to an O_2 plasma for the following sets of data: (a) Data derived from a photo-enlargement of Figure 1 of Morra and co-workers (1) with the aid of a variable scale for accurate interpolations; (b) prior ESCA data from Golub and co-workers (2) for *very clean* PTFE, showing no evidence for hydrocarbon contamination (initial F/C = 2.0), and exposed for 10, 20 and 30 min 'in the glow' of an O_2 plasma, yielding oxygen uptakes of 0.11, 0.15 and 0.21 atom % O, respectively; and (c) new data obtained expressly for the recent article by the latter workers (3), using disks cut from a 25-mm thick PTFE sheet (initial F/C = 1.96, O/C = 0.0098) similar to that used previously (2) and having a small amount of hydrocarbon contamination, but much less than in the PTFE sample used by Morra and co-workers. Figure 3 shows the C_{1s} ESCA spectra of this 'new' PTFE before and after various

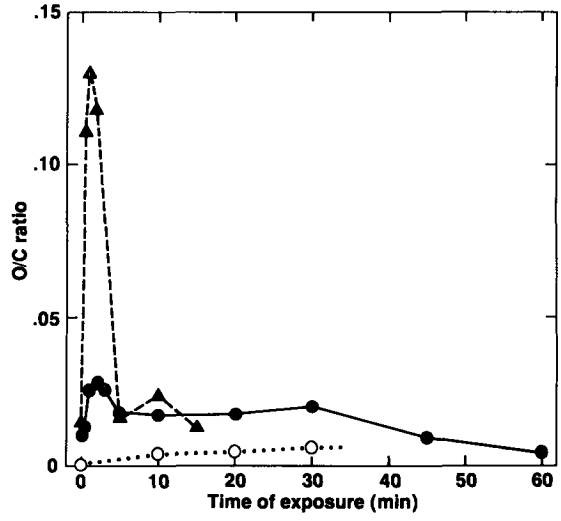


Figure 1. Effect of time of exposure to O_2 plasma on the O/C ratio of PTFE, as determined from ESCA spectra: ●, new data; O, prior data from Golub and co-workers (2); ▲, data from Morra and co-workers (1).

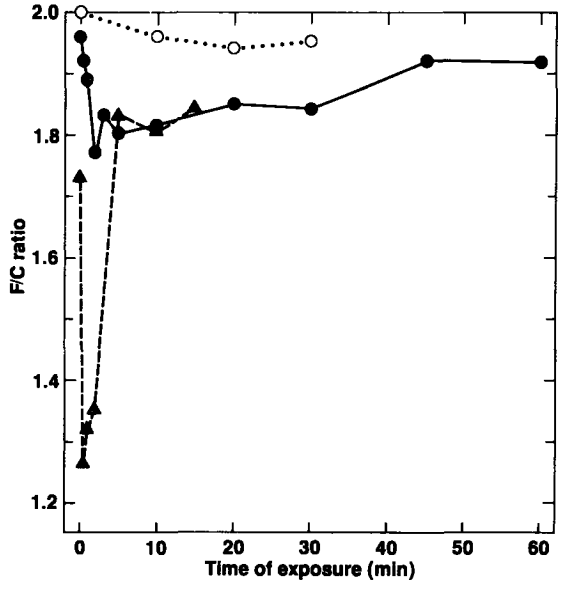


Figure 2. Effect of time of exposure to O_2 plasma on the F/C ratio of PTFE, as determined from ESCA spectra: ●, new data; O, prior data from Golub and co-workers (2); ▲, data from Morra and co-workers (1).

times of exposure in the same O_2 plasma reactor (and obtained with the same SSX-101 ESCA spectrometer) used to obtain the prior data for the very clean PTFE mentioned under data set (b) above. A crude measure of the hydrocarbon contamination in the 'new' PTFE is given by the ratio of intensities of the peaks at ca. 285 eV (-CH_2- and/or $\dot{C}H$-) and 292 eV (-CF_2-); for the unexposed film, this ratio or II/I is 0.027. In contrast, the corresponding ratio for the unexposed PTFE used by Morra and co-workers is much larger, which we determined as 0.15 by planimetering the areas under the two peaks in a photoenlargement of their Figure 2.

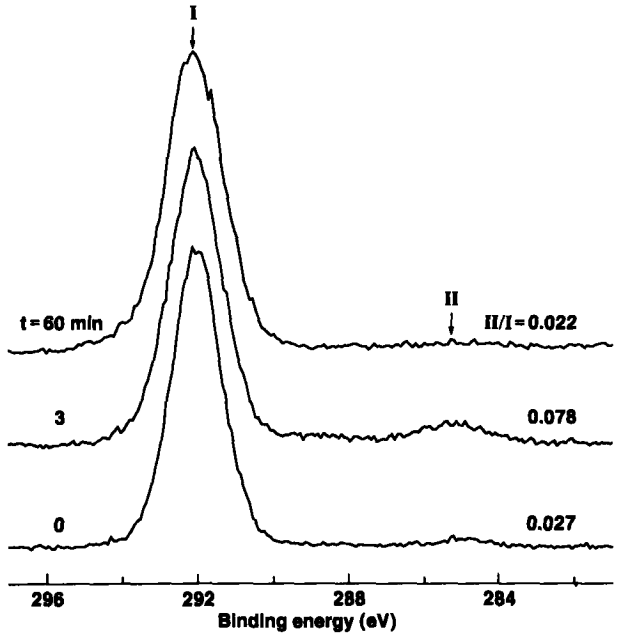


Figure 3. C_{1s} ESCA spectra of 'new' PTFE before and after exposure to O_2 plasma for 3 and 60 min. Peak I centered at ca. 292 eV corresponds to -CF_2- groups, but may overlap contribution from $\dot{C}F$- groups, while peak II at ca. 285 eV corresponds to contributions from hydrocarbon contamination (-CH_2-, $\dot{C}H$-) and oxidized carbon ($\geq$C-O-, $>$C=O).

A better measure of the hydrocarbon content in the unexposed PTFE takes into account the fact that, for small amounts of oxidation, peaks due to $\geq C$-O- or $>C$=O (typically located at 286.0-286.4 eV) will merge with that for -CH_2- / $\dot{C}H$- groups (284.6 eV) into the observed broad 285-eV peak. Also, for the O_2-plasma treated PTFE, where small amounts of $\dot{C}F$- groups may arise from fluorine detachment, the ESCA peak of such groups (at 290.4 eV) will be concealed under the low-eV tail of the broad 292-eV peak, which is due principally to -CF_2- groups (centered at 291.9 eV). (Support for the latter point is seen in the ESCA spectra of Teflon FEP, a tetrafluoroethylene-hexafluoropropylene copolymer (2)) Thus, II/I is a measure of the following ratio of carbon atoms: [-CH_2-, $\dot{C}H$-, $\geq$C-O-, $>$C=O]/[-CF_2-, $\dot{C}F$-]. For simplicity, we will refer to -CH_2- and/or -CH- as CH, $\geq$C-O- and/or $>$ C=O as CO, and -CF_2- and $\dot{C}F$- as CF_2 and CF, respectively. Now, noting that F/C = 1.96 and O/C = 0.0098 for the unexposed PTFE in Figure 3, we infer that the initial 'new' PTFE comprises 98 CF_2, 1 CH and 1 CO for every 100 carbon atoms, for an oxygen level of 0.34 atom % O and a calculated ratio for II/I = 0.020; this ratio compares reasonably well with the ESCA-determined ratio of II/I = 0.027. In a similar way, we infer that the unexposed PTFE used by Morra and co-workers (F/C = 1.73; O/C = 0.014) comprises 86.5 CF_2, 12.1 CH and 1.4 CO for every 100 carbon atoms, for an oxygen level of 0.51 atom % O and a calculated ratio for II/I of 0.16, which is virtually the same as the planimeter-derived ratio of 0.15 indicated above.

Returning to the present Figure 1, we observe that Morra's PTFE, with a hydrocarbon content (or CH) ca. 12.1 times that of the 'new' PTFE, exhibits a spike in the O/C-exposure time plot (or Δ(O/C)) whose amplitude is ca. 0.115/0.017, or 6.7, times the amplitude of the spike for the 'new' PTFE. For the *very clean* PTFE (of our prior data (2)), which has no detectable hydrocarbon contamination, there is no spike at all, only a

small, monotonic increase of O/C ratio with time of exposure, tending towards an "equilibrium" oxygen uptake after prolonged exposure as a result of the dynamic competition between oxidation and etching (or surface regeneration). To the extent that there is oxygen uptake by the O_2 plasma-exposed PTFE, the F/C ratio necessarily decreases, whether or not fluorine detachment occurs. This is reflected in Figure 2, which shows changes in the F/C ratios accompanying the changes in the O/C ratios presented in Figure 1. As in the latter figure, we see in Figure 2 that the amplitude of the spike in the F/C-exposure time plot for Morra's PTFE (Δ(F/C) = -0.47) is greater than that (-0.19) for the 'new' PTFE, while the very clean PTFE again shows no such spike. Apropos, it is worth recalling our earlier observation (2) that "apparently there is a correlation between the [oxygen] uptake and the level of hydrocarbon contamination for [PTFE] films exposed to [oxygen atoms in low Earth orbit (LEO)] on the [Space Shuttle] STS-8, and we may speculate that the $-CF_2CF_2-$ structure per se undergoes negligible oxygen uptake on exposure to $O(^3P)$ in LEO." Here we wish to stress the need for working with very clean PTFE because the surface properties of this polymer, which has great industrial importance and is easily contaminated by ubiquitous hydrocarbon, are therefore often difficult to reproduce. At any rate, it follows that the changes in surface properties (namely, wetting behavior and contact angle hysteresis) reported *at short treatment time* by Morra and co-workers cannot be considered representative of pristine PTFE.

That a steady-state surface composition is approached on prolonged exposure to O_2 plasma is seen not only in Figures 1 and 2 but also in Figure 3, where the II/I ratio increases from an initial 0.027 to 0.078 for the 3-min exposure and subsequently drops to 0.022 for the 60-min exposure. Given the II/I ratio and the O/C and F/C ratios (0.026 and 1.83, respectively) for the 'new' PTFE exposed for 3 min, we calculate that its structure comprises 90.0 CF_2, 2.8 CF, 4.6 CH and 2.6 CO for every 100 carbon atoms, for an oxygen level of 0.91 atom % O. Again, given the II/I ratio and the O/C and F/C ratios (0.0038 and 1.92, respectively) for the 'new' PTFE exposed for 60 min, *its* structure is presumed to comprise 94.0 CF_2, 3.8 CF, 1.8 CH and 0.4 CO for every 100 carbon atoms, for an oxygen level of 0.14 atom % O. The latter value, which is in the range observed earlier for a pristine PTFE exposed to an O_2 plasma for 10-30 min, is less than the initial value (0.34 atom % O) of the unexposed 'new' PTFE.

A similar approach towards a steady-state composition may be deduced from the ESCA results of Morra and co-workers. From Figure 2 of ref 1, the II/I ratio for their PTFE is 0.15 initially and 0.12 after a 15-min exposure to an O_2 plasma. At intermediate exposures, the II/I ratio apparently increases to ca. 0.48 after 1-min exposure, then decreases to ca. 0.070 after 5-min exposure, and remains at ca. 0.087 thereafter. This last statement is seen as follows: Given the O/C and F/C ratios (0.129 and 1.32, respectively) for Morra's PTFE exposed for 1 min, a plausible structure comprises 64.5 CF_2, 3.0 CF, 19.6 CH and 12.9 CO for every 100 carbon atoms. Such a structure yields the ratio II/I = 0.48 just mentioned and, significantly, an oxygen level of 5.3 atom % O -- substantially higher than that encountered in any PTFE exposed to oxygen atoms, including two samples recovered from the STS-8 mission that exhibited some hydrocarbon contamination (2). Again, given the O/C and F/C ratios (0.016 and 0.083, respectively) for Morra's PTFE exposed for 5 min, a plausible structure is 89.0 CF_2, 4.5 CF, 4.9 CH and 1.6 CO, which yields a II/I ratio of 0.070 and an oxygen level of 0.56 atom % O. Lastly, the PTFE exposed for 15 min, given the pertinent O/C and F/C ratios (0.0125 and 1.84, respectively), has a structure (92 CF_2, 6.7 CH, 1.3 CO) which yields a II/I ratio of 0.087, somewhat less than the 0.12 measured planimetrically, and an oxygen level of 0.46 atom % O. To interpret the ESCA results of Morra and co-workers, it is clear that the hydrocarbon present in the PTFE surface is preferentially oxidized in the O_2 plasma, and the oxygen uptake can rapidly reach high levels; for polyethylene exposed to O atoms (4), for example, the "saturation" level is ca. 15-22 atom % O. For Morra's PTFE sample exposed for 1 min, and with a hydrocarbon content (CH/(CF_2 + CF + CH)) of ca. 23%, an oxygen uptake of 0.23 × (15-22) or 3.5-5.1 atom % O may be anticipated; as we saw above, the oxygen level for that sample (5.3 atom % O) is indeed close to this range. Upon prolonged exposure, of course, the hydrocarbon should be fully oxidized away, and this accounts for the fact that Morra's PTFE surface is then comparable to, or even cleaner than, that of the unexposed polymer. On a smaller scale, because of its lower hydrocarbon content, the 'new' PTFE results present the same picture.

There is one additional feature of the foregoing results worth mentioning. Both for Morra's PTFE and the 'new' PTFE, there are definite ESCA indications of an increase in hydrocarbon contamination at short times of exposure, followed by a decrease to the level of the initial polymer, or even below it, upon prolonged exposure. Thus, the CH content in Morra's PTFE increased from ca. 12.1 (per 100 carbon atoms) to ca. 19.6 (at 1-min exposure), decreasing to ca. 4.9 (at 5-min exposure), and ending with ca. 6.7 (at 15-min exposure). Similarly, the CH content in the 'new' PTFE increased from 1.0 (per 100 carbon atoms) to 4.6 (at 3 min-exposure) and decreased

thereafter to 1.8 (at 60-min exposure). This trend of an initial increase, followed by a decrease, in hydrocarbon content with time of exposure is probably not an artifact from handling exposed PTFE samples. Instead, it suggests that as the hydrocarbon content in the surface is oxidized away, additional hydrocarbon "blooms" to the surface of the PTFE (in the manner of blooming of compounding ingredients in vulcanized rubber formulations) from underlying molecular layers until the hydrocarbon in the bulk is eventually removed through the etching process.

ACKNOWLEDGMENT

The authors thank Plasma Science, Inc., then of Belmont, now of Foster City, California, for performing the new set of exposures of PTFE to an O_2 plasma discussed in this paper.

REFERENCES

(1) Morra, M., Occhiello, E., and Garbassi, F.: Langmuir vol 5, 1989, p. 872.
(2) Golub, M. A., Wydeven, T., and Cormia, R. D.: Polymer vol 30, 1989, p. 1571.
(2) Golub, M. A., Wydeven, T. and Cormia, R. D.: Langmuir vol 7, 1991, p. 1026.
(4) Golub, M. A., and Cormia, R.D.: Polymer vol 30, 1989, p. 1576.

MEDICAL ADVANCES

(Session E6/Room C4)

Thursday December 5, 1991

- **Applications of the Strategic Defense Initiative's Compact Accelerator Technology**
- **Acoustically-Based Fetal Heart Rate Monitor**
- **Surgical Force Detection Probe**
- **Dynamic Inter-Limb Resistance Exercise Device**

PRECEDING PAGE BLANK NOT FILMED

N92-22732

APPLICATIONS OF THE STRATEGIC DEFENSE INITIATIVE'S COMPACT ACCELERATORS

Nick Montanarelli and Ted Lynch[1]
Strategic Defense Initiative Organization
Office of Technology Applications
The Pentagon
Washington, DC 20301-7100

ABSTRACT

The Strategic Defense Initiative's (SDI) investment in particle accelerator technology for its directed energy weapons program has produced breakthroughs in the size and power of new accelerators. These accelerators, in turn, have produced spinoffs in several areas: the radio frequency quadrupole linear accelerator (RFQ linac) was recently incorporated into the design of a cancer therapy unit at the Loma Linda University Medical Center, an SDI-sponsored compact induction linear accelerator may replace Cobalt-60 radiation and hazardous ethylene-oxide as a method for sterilizing medical products, and other SDIO-funded accelerators may be used to produce the radioactive isotopes oxygen 15, nitrogen 13, carbon 11, and fluorine 18 for positron emission tomography (PET). Other applications of these accelerators include bomb detection, non-destructive inspection, decomposing toxic substances in contaminated ground water, and eliminating nuclear waste.

INTRODUCTION

Particle accelerators, devices that produce high-energy beams of charged atomic or sub-atomic particles, have largely been limited to research applications due to their high cost. SDI, however, has focused much attention on developing low-cost, reliable particle accelerators as part of a system to provide protection against ballistic missile attacks. As a result, several SDI-funded researchers are developing ways to reduce the size, weight, and cost and increase the reliability of particle accelerators that drive free electron lasers, neutral particle beams, and other directed energy weapons. As a result of these improvements, SDI-funded accelerators have a variety of spinoff applications.

Researchers have long known accelerator technology could be used for a number of medical and industrial applications such as providing treatments for cancer and other ailments, sterilizing medical products, production of isotopes for PET imaging, non-destructive inspection and testing, industrial welding, environmental clean-up, and electron-beam processing. Widespread application has never been achieved, however, due to limitations in accelerator technology that prevented them from replacing alternative techniques, such as employing radioactive sources. Thus, the same improvements in size, weight, and cost sought by SDI would benefit these commercial applications. As a result, several accelerators developed with SDI funding have made these applications a near-term reality instead of the dream of a few research accelerator physicists.

---

[1] Nick Montanarelli is Deputy Director, Office of Technology Applications, for SDIO's Innovative Science and Technology Directorate. Ted Lynch is a technical writer with Systems Engineering and Management Associates (SEMA), Inc., a technology management contractor for the Office of Technology Applications.

Our thanks to Dr. Joseph Mangano, Science Research Laboratory, Inc.; Dr. Robert W. Hamm, AccSys Technology, Inc.; Dr. William Hagan, Science Applications International Corporation; and John R. Gustafson, Los Alamos National Laboratory for their cooperation in preparing this report.

PRECEDING PAGE BLANK NOT FILMED

APPLICATIONS

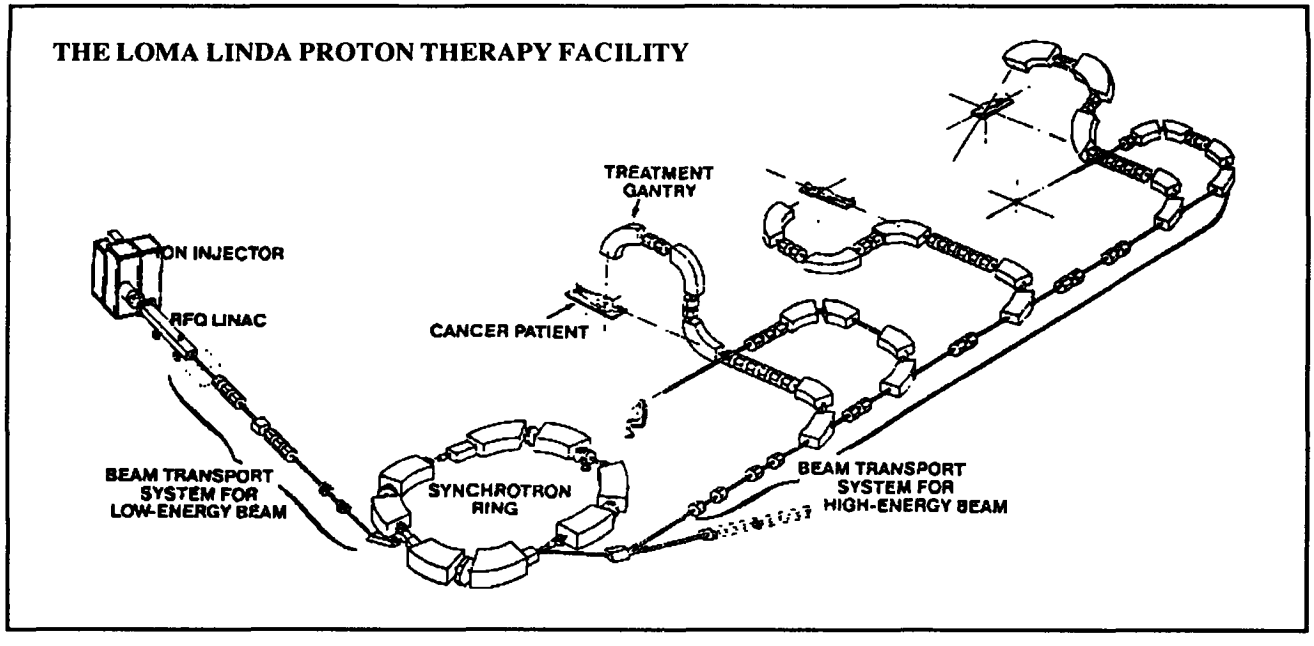
Proton Cancer Treatment at Loma Linda

The radio frequency quardupole linear accelerator (RFQ linac) — a key component of SDI's neutral particle beam program — now serves as the first stage of a high-energy proton accelerator operating at the Loma Linda University Medical Center to treat cancer.[2] Proton cancer therapy can treat a wide range of tumors — from those in the digestive system to those in the eye and brain. In addition, proton therapy is safer than conventional radiation or chemotherapy. The second two techniques both kill cancerous cells; however, because therapists have very little control over where these treatments are deposited, they are also more likely to damage adjacent healthy tissue. In fact, for some deep tumors, the treatment attacks more intervening healthy tissue than the tumor itself. As a result, the therapist often lowers the dose to the tumor to prevent excessive damage to surrounding tissue. This conservative approach may allow the tumor to continue to grow.

In contrast, therapists can precisely control proton beams. This is because proton beams deposit nearly all their energy at the target with little scattering. Also, since protons are charged particles, the therapist can more precisely control a proton beam's path than other forms of commonly used radiation (e.g. x-rays, gamma rays). In addition, protons produce much less damage to adjacent healthy tissue, making it possible to administer protons in higher doses. As a result, this treatment more thoroughly destroys the tumor with less side effects.

Other proton threapy facilities are also planned for Massachusetts General Hospital and the University of California at Berkeley.

Figure 1. The Loma Linda Proton Therapy Facility



Sterilization of Medical Products

Technologists have studied electron-beam conversion to x-ray radiation as a method to sterilize medical products for the last three decades. This method, however, has not been economically feasible, since the accelerator technology required to produce the electron-beams has never been competitive with chemical and radioisotopic methods of sterilization. In recent years, though, electron-beam sterilization has become a more viable alternative. Chemical sterilization using

[2] AccSys Technology, Inc. provided the RFQ linac injector for this facility, while Science Applications International Corporation (SAIC) installed the entire synchrotron system. Both companies have done extensive work on accelerators for SDI which contributed to this project.

ethylene oxide (ETO), is losing favor because ETO is an extremely explosive gas that must be stabilized with a buffer of chloro-fluorocarbons (CFCs). Because CFCs are associated with destruction of the ozone layer, recent legislation has placed a 100 percent tax on use of ETO gas mixtures and mandated almost complete recovery of CFCs. In addition, the EPA has strongly discouraged use of ETOs and perhaps may ban all but essential use of ETOs within the next decade.

Radioisotopic sterilization uses Cobalt 60 to produce x-ray radiation to sterilize medical products. Nordian International, a crown corporation of the Canadian Government, supplies 80 percent of all Cobalt 60, which is produced by neutron absorption of Cobalt 59 in nuclear reactors. The cost of Cobalt 60 has gone up over 50 percent in two years after the Canadian Government privatized the supply of Cobalt 60. In addition, Cobalt 60 sterilization plants face increasing costs associated with the storage, handling, and disposal of Cobalt 60.

Because electron-beam irradiation is forward-directed, it could sterilize medical products on a conveyor line instead of on a pallet, as is now done with ETO and Cobalt 60. Because a conveyor line system could be integrated with other production lines, it would further decrease costs and provide greater assurance of quality control.

Even with these advantages, however, the electrostatic and RF accelerator technologies available to date are not likely to supplant Cobalt 60 sterilization, due to their high operating costs. Science Research Laboratories (SRL), Inc., however, has developed a compact, modular linear accelerator with SDIO support that could break into this market. SRL's SNOMAD IV accelerator has a lower capital cost than competing accelerators and Cobalt 60; in addition, it also has much lower operating costs due to its reliability.

Production of PET Isotopes

In 1989, SDIO started the miniaturized Positron Emission Tomography (PET) accelerator program to reduce the cost of producing certain elements necessary for PET imaging. PET is important for medical applications because it images the body's chemical processes. It works as a diagnostic tool for cancer, brain disease, heart disease, and as an important research tool to enhance our understanding of the brain and mental disorders. The SDIO PET program builds on its large investment in accelerator technology for the SDIO neutral particle beam (NPB) program and will benefit both PET and SDIO, since PET and the NPB require similar accelerator technologies and the same large-scale manufacturing.

Currently, use of PET is largely limited to research uses because of the high cost of producing the tracer elements used for medical imaging. Large, high-power cyclotrons are used to produce these radiopharmaceuticals; while appropriate for research applications, where a wide variety of radiopharmaceuticals are needed in a limited supply, cyclotrons are extremely uneconomical for widespread clinical use. Thus, development of compact, inexpensive accelerators is generally considered the ideal way to introduce widespread clinical use of PET.

Accelerators now being developed for clinical uses have much lower power requirements than cyclotrons but still produce the four most common radiopharmaceuticals — fluorine 18, nitrogen 13, oxygen 15 and carbon 11. To get adequate quantities of the radiopharmaceuticals using low-power levels, the accelerators must produce high-current beams. SDIO is currently funding development of two small, low-power, high-current accelerators for the PET program: an RFQ system designed by SAIC and an electrostatic accelerator designed by Science Research Laboratories, Inc.

Other applications (electron-beam)

SDIO has sponsored the development of several compact electron-beam accelerators to drive free electron lasers. Non-medical applications of these accelerators are listed here.

- Coal-fired power plants can use electron-beams to reduce emissions of sulphur and nitrogen oxide, which cause acid rain, by converting them into the common fertilizers ammonium sulfate and ammonium nitrate. In addition to reducing acid rain, this method would allow power plants to use high sulphur-content coal. High-sulfur coal is the nation's most abundant fossil fuel; thus, this technique would greatly reduce our dependence on foreign oil without harming the environment.

- X-rays generated by electron-beams can treat ground and waste water contaminated with toxic substances. Most toxic wastes are artificial molecules that do not occur in nature. Electron-beams break these molecules into fragments, which afterward tend to recombine into simpler, non-toxic substances. The radiation dose required to accomplish this depends on the particular toxic material and the substance mixed with it. Dose requirements range from a small fraction of a Mrad (a measure of the amount of radiation absorbed) to over 10 Mrads, easily in the range of several SDI-developed accelerators.

- Electron-beam-generated x-rays can irradiate meats, fruits, vegetables and other perishable foods to prevent them from spoiling. This method provides a safer, non-nuclear source of radiation and could eliminate the need to treat food with potentially harmful chemicals.

- Electron-beams could treat the surface of materials. The technique can be used to harden plastics and other materials, improve the heat resistivity of wire insulation, control the quality of automobile rubber tires, and cure paintings or printing inks. It can also be used to bond metal-matrix composites, cross-link plastics and join ceramics.

- High energy electron-beams (on the order of 10 MeV) could be used to weld several types of materials, including HY100 (a high-strength steel used primarily in submarine hulls), aluminum, stainless steel, and titanium. Electron-beam welding has several advantages over other techniques:

  - The process can be done at atmospheric pressure. In other processes, parts need to be welded in a vacuum chamber, This makes high-energy electron-beam welding especially attractive for welding aircraft carrier deckplates, submarine hulls, nuclear power plant facilities, and other large-scale construction projects.

  - By penetrating further into the material at high energy densities, electron-beams create stronger, deeper welds. Electron-beams heat an area 5 mm into the material, while normal welding processes only heat the material's surface. As a result, normal welding processes cause heat stresses that weaken the material.

  - The radiation processes allow you to see the weld as the electron-beam forms it, resulting in better quality control.

Other applications (neutrons)

As part of the neutral particle beam program, SDIO has sponsored the development of particle accelerators that produce beams of neutrons by accelerating protons against a metal target. The interaction between the protons and the metal target produces a stream of neutrons that is used for these non-medical applications.

- The Federal Aviation Administration has developed a lightweight bomb detector that uses AccSys Technology, Inc.'s RFQ linac as a neutron source (SRL and SAIC have developed neutron sources for this purpose, as well: the SRL Tandem Cascade Accelerator [TCA] and SAIC's RFQ linac). The FAA bomb detector, which is currently undergoing testing, bathes luggage in low-energy neutrons. High-nitrogen-content explosives that may be in the luggage absorb the neutrons and emit a characteristic gamma radiation that can be detected by sensors in the system. Unlike other portable sources of neutrons which employ radioactive materials, the RFQ linac avoids radiation hazards and is lighter, because bulk shielding is not required. The compact size will allow FAA to employ bomb detectors in more airports, and give the airports greater flexibility of use, since the bomb detector can be more easily moved from terminal to terminal.

- The RFQ Linac also can provide neutrons for neutron radiography. This technique is used to detect elements that selectively absorb neutrons, including hydrocarbons (found in oil and O-ring seals) and hydroxides (found in corroded aluminum). Thus, it can inspect airplanes for cracks or corrosion, artillery shells for cracked explosive charges, and rocket engine propellants during test firing. It can also determine lubricant flow in aircraft engines and detect oil deposits.

- SDI accelerators could also change long-lived nuclear wastes into harmless isotopes (see figure 2). A process under development at Los Alamos National Laboratory would use a high-power, high-current accelerator to create an intense flux of neutrons. A molten salt would carry the nuclear waste through a tank of heavy water (water made up of the heavier hydrogen isotope deuterium), which slows down the neutron flux.

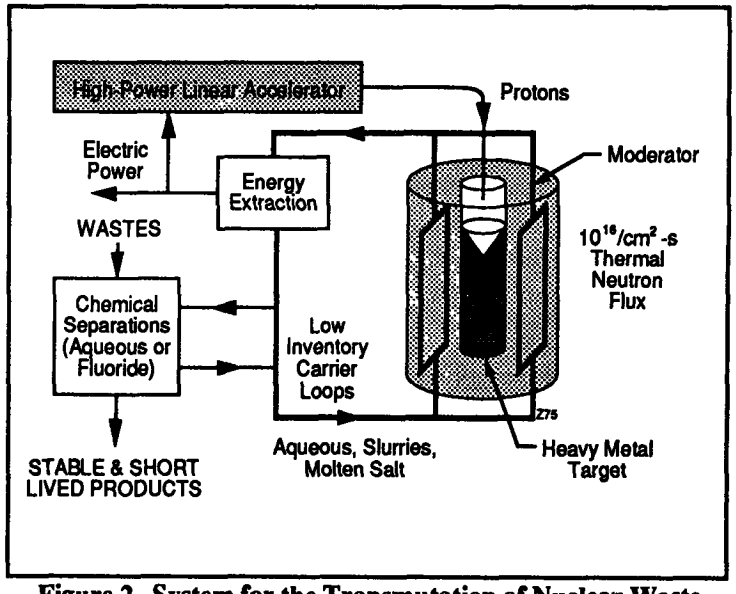


Figure 2. System for the Transmutation of Nuclear Waste

By slowing down the neutrons, the heavy water allows more nuclear waste to absorb the neutrons; the neutrons, in turn, spark a nuclear reaction that changes the waste into a new, stable isotope or into an unstable isotope that will quickly decay into a stable one. The molten salt, which has been heated up during this reaction, is then circulated out of the tank of heavy water. Once out of the heavy water, the molten salt can be used to generate electricity. Then, after the process is complete, chemical reactions extract the harmless isotopes out of the molten salt and replace it with more nuclear waste for another round of processing.

ACCELERATORS

Radio Frequency Quadrupole Linear Accelerator (RFQ Linac)

Soviet accelerator physicists developed the first RFQ linac in 1974. Since learning of this development in 1977, Western scientists have used this concept to replace the three-story-high electrostatic accelerator (current RFQ linacs are 3-10 feet long and 1-3 feet in diameter). The RFQ linac's body consists of a quadrupole cavity, with 4 electrode vanes that protrude from the cavity walls in the shape of a plus (+) sign. This cavity is designed to produce oscillating electric fields between the tips of the electrodes when radio frequency power is applied to it. The oscillating fields bunch, focus, and accelerate ions or protons, providing a high-quality, high-current beam.

RFQ linacs were initially developed to serve as the first stage of high-energy linacs used for physics research (and later for SDI neutral particle beams, cancer therapy facilities like Loma Linda, and the production of tritium). High-power linacs such as these require high-current ion beams accelerated to energies of about 1 MeV. Prior to the invention of RFQ linacs, very large electrostatic accelerators, which require complex beam-focusing systems, served this purpose.

An RFQ linac now serves as the initial stage in over 20 research facilities, including the European Center for Nuclear Research (CERN) facility near Geneva Switzerland. In addition, the RFQ linac will play an important role in the U.S. Superconducting Super Collider, both as the initial stage injector, and to calibrate many of its particle detectors. For other, low-power applications, such as those mentioned earlier, the RFQ linac can operate as a stand-alone system.

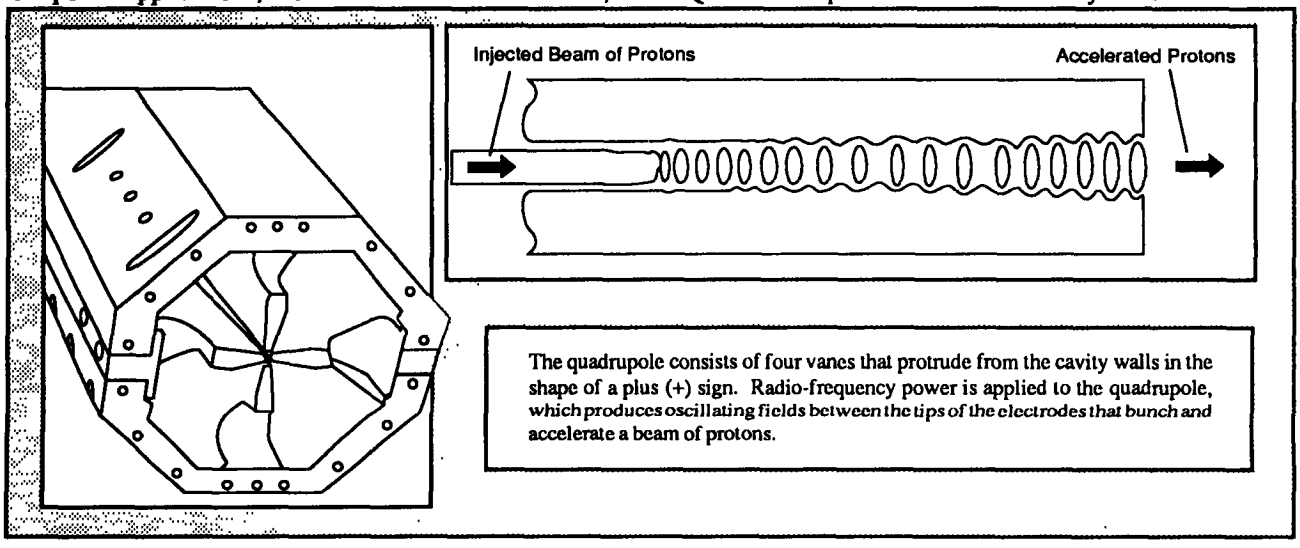


Figure 3. The Four-Vaned Configuration of the RFQ Linac

Science Research Laboratory's SNOMAD IV

SRL developed the SNOMAD IV as part of an SDIO Small Business Innovation Research project to build an electron-beam driver for the free electron laser. A key feature of the SNOMAD IV is the all solid-state driver, which dramatically increases reliability; the SNOMAD IV can operate for 2-3 years (generating 10^{11} to 10^{12} shots) without maintenance, unlike other accelerators which must be serviced every few months. Also, since the SNOMAD IV would use six accelerator modules to produce the electron-beam energies necessary for bulk sterilization, mechanical breakdowns could be serviced with minimal down-time. A spare module could simply replace the affected module while someone serviced it. Finally, the SNOMAD IV has a completely computerized control panel that allows a technician to punch in the dose and run-time. This

ease-of-use reduces labor costs by eliminating the need for a staff of highly trained accelerator physicists.

The SNOMAD IV also produces much higher average currents than competing RF and electrostatic accelerators. The higher currents allow the accelerator to operate at optimal beam energies (8 MeV) without sacrificing throughput (average power). Beam energies much higher than 8 MeV are undesirable because medical products, when subjected to beam energies greater than 10 MeV, can become permanently radioactive. Because other accelerators have low average currents, they must sacrifice throughput to operate below this 10 MeV maximum.

Table 1. SRL's SNOMAD IV Compared to Competing Electron-Beam Accelerators

Characteristics	SNOMAD IV	RF Accelerators (AECL's IMPELA)	Electrostatic Accelerators (DYNAMATRON)
Average Power* (Throughput)	1 Megawatt	50 kilowatts	200 kilowatts
Beam Energy (measures penetration)	8 MeV	10 MeV	5 MeV
Average Current	0.125 Amps	0.005 Amps	0.04 Amps
Capital Cost	$2.5 million ($400,000/module)	$3.5 - 4 million	$1.5 - $4 million

* note: average power = beam energy x average current

Miniaturized Accelerators for PET

The SDIO-funded PET accelerators have been designed for easy maintenance, easy operation and low shielding requirements — all of which reduce operating costs for clinical PET radiopharmaceutical production. Cyclotron systems have long maintenance down-times — often forcing an operating schedule of four days on and three days off — because there is a long "cool-down" period for the radioactive materials before maintenance can be performed. The SDIO-funded accelerators have little — if any — cool-down period and less maintenance requirements in general. Also, a staff of accelerator physicists must operate cyclotrons, while a technician working at a computer terminal could operate the SDIO-funded accelerators.

Finally, the shielding requirements for the SDIO-funded accelerators are much lower. Because the SAIC RFQ linac accelerates Helium-3 particles instead of protons or deuterons, it has shielding requirements ten times lower than for cyclotrons. This is because Helium 3 particles produce very few neutrons when interacting with the target material, and therefore produce less hazardous radiation.

Table 2. Comparison of Accelerators for PET Radiopharmaceutical Production

PET Accelerators	SAIC RFQ design	SRL Tandem Cascade Accelerator (TCA)	Cyclotrons
Average Power	20 kW	37 kW	High
Beam Energy	8 MeV (accelerating He^3)	3.7 MeV (accelerating protons and deuterons)	Varies (depending on current produced) — up to 50-60 MeV
Average Current	300 microamps	1 mA	Varies
Capital Cost	under development (less than $1 million)	$750,000	$1-2 million
Weight	1,300 lbs.	1,200 lbs.	5-20 tons
Radiopharmaceuticals produced	all (^{11}C, ^{15}O, ^{18}F, ^{13}N)	all	all (plus others for research purposes)

Transmutation of Nuclear Waste Accelerator

Transmutation accelerators require extremely high-power proton-beams to produce adequate neutron-fluxes to change nuclear waste into stable by-products. The Los Alamos design — originally developed for tritium production (which requires much higher currents but similar power levels) — uses two pairs of injectors, each consisting of two RFQ linacs and two drift-tube linacs (DTL). The injector-produced proton-beams are combined into a 20 MeV beam using a funneled beam-launcher (see diagram). The 20 MeV beam is then accelerated through a coupled-cavity linac, producing a 1600 MeV beam with a current of 50-250 mA (55 mA for transmutation and 250 mA for tritium production).

These stages are designed to optimize particle acceleration at successfully higher velocities. Advances in high-current linear accelerator technology initiated by SDI's Neutral Particle Beam program have produced sizable improvements in the generation, acceleration, and handling of low-energy beams within the accelerator. Because efficient low-energy handling during the injector phase is key to creating a reliable, high-current, high-energy accelerator system, these advances have made the transmutation system possible.

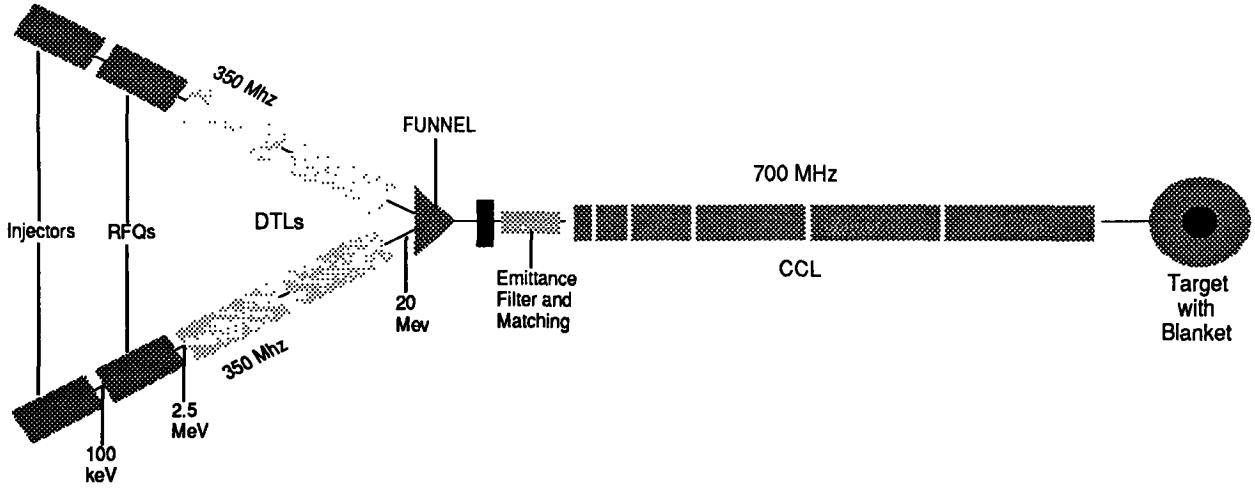


Figure 4. Proposed Accelerator for the Transmutation of Nuclear Waste

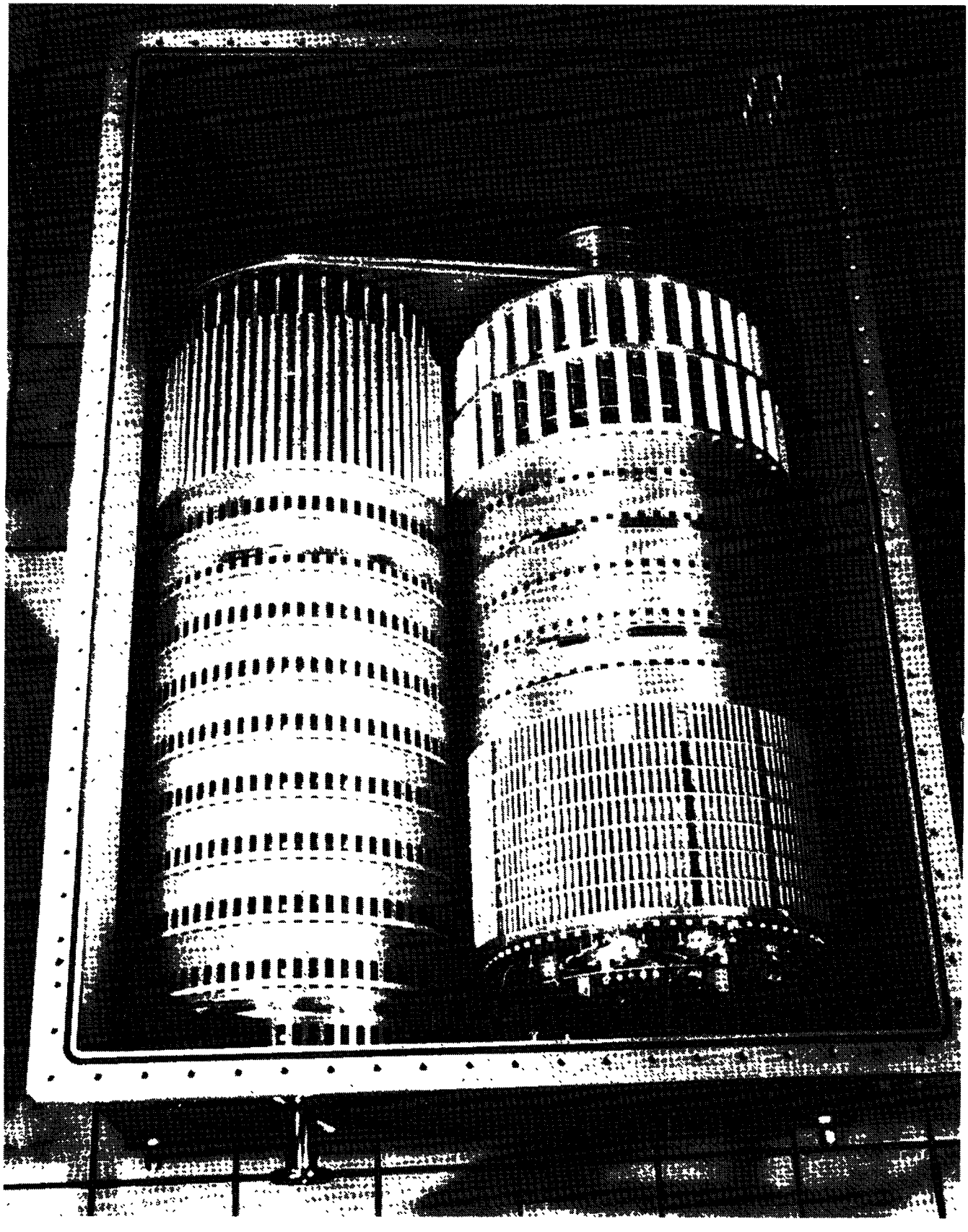


Photo #1: One Module of SRL's SNOMAD IV Accelerator

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Photo #2: SRL's Tandem Cascade Accelerator

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

N92-22733

ACOUSTICALLY BASED FETAL HEART RATE MONITOR

Donald A. Baker, MD
Baker Guardian Medical Labs
Spokane, WA 99207

Allan J. Zuckerwar
NASA Langley Research Center
Hampton, VA 23665

ABSTRACT

The acoustically based fetal heart rate monitor permits an expectant mother to perform the fetal Non-Stress Test in her home. The potential market would include the one million U.S. pregnancies per year requiring this type of prenatal surveillance. The monitor uses PVF2 piezoelectric polymer film for the acoustic sensors, which are mounted in a seven-element array on a cummerbund. Evaluation of the sensor output signals utilizes a digital signal processor, which performs a linear prediction routine in real time. Clinical tests reveal that the acoustically based monitor provides Non-Stress Test records which are comparable to those obtained with a commercial ultrasonic transducer.

INTRODUCTION

It has been known for a long time that fetal heart rate (FHR) patterns reflect fetal well-being. If a fetus's oxygenation status deteriorates (early asphyxiation), its heart rate will compensatorily increase. This is not surprising because adult hearts respond similarly when adults get short of breath (a form of asphyxiation) with exercise. Later, with continued hypoxic *stress*, the FHR slows due to metabolic stagnation and creates a true threat to fetal well-being termed *fetal distress*. The FHR, therefore, would be expected to accelerate with in-utero fetal movement. This is what happens in the healthy fetus while the absence of FHR accelerations suggests that the fetus may be in jeopardy. This knowledge has sponsored the impetus for generating our current hospital-based FHR monitoring equipment and establishing FHR reference standards. The best known test of fetal well-being is the Non-Stress Test (NST), which relies on FHR changes in relationship to fetal movement.

PORTABLIZING THE NON-STRESS TEST

Each NST costs anywhere from $125-250, entails hospital paperwork, inconvenience and fragmentation of care for the patient, and requires the attendance of nursing personnel who are presently in short supply. All of this technology could be converted to a portable computer-based, automated form utilizing a plurality of passive surface pressure sensors capable of performing the same testing. Such an approach would not involve the intentional creation and injection of an energy source into the pregnant woman and her fetus as occurs with ultrasound. Such automation would be more economical, thus improving the potential for increased fetal surveillance while easing the burden on strained nursing resources. Further, automation would be more scientific than its error-prone subjective strip chart interpretation counterpart. Its portability feature could overcome certain patient limitations (rural areas, patient paralysis, complete lack of transportation...) and in some areas the cultural paranoia of the "white coat" community by making itself available in the home or a local satellite facility. Further, by increasing prenatal surveillance, tragic outcomes such as cerebral palsy could be prevented one day.

The initial market would target the obstetrician's office and his high-risk Ob population. He would employ it and would receive direct payment instead of the hospital. The revenue derived would enable him to purchase the monitor rather quickly while appealing to his patients' sense of economy and convenience. Ultimately the patient would apply the passive sensor array/abdominal cummerbund unit to herself at home.

Either data transmission via phone link or an on-site automatic readout would be available in that case. In all circumstances, the commonly accepted NST would be performed.

The NST is routinely covered by insurance companies and public assistance facilities. Since one million of our three million annual pregnancies require periodic NST evaluations, realistic commercialization forecasts have predicted a $100 million per year industry potential in the U.S. alone for portable fetal monitoring. This prediction does not include the potential service revenue to distributors nor its tangible or intangible preventative health care value to society in general.

The previously mentioned automatic nature of the unit would lend itself to patient or para-professional application. In part, the automatic feature relies on a plurality of lightweight surface pressure sensors attached to an attractive cummerbund which the mother places on her abdomen. This is needed because the characteristic fetal heart sound is a localized, yet randomly placed acoustic event. The fetal heart sound is heard only in proximity to where the fetal back is in contact with the maternal abdomen--a roughly circular area having a diameter of about 3 cm. Since the fetus moves with some frequency, a plurality of transducers is required to anticipate all locations of this expected sound. Coupled to this plurality is a computer programmed to scan over the transducers with fetal heart sound and rate recognition capabilities. Either the mother or an accelerometer activation would signify when fetal movement occurred. Once tracking the FHR/movement events, the computer would automatically compare the tested fetal NST to acceptable norms, the outcome of which would be relayed to the physician via automatic strip chart or data phone link. Such NST could be performed at $25/test--a 10-fold saving over our current expenses.

TECHNICAL DESCRIPTION OF THE MONITOR

A block diagram of the monitor hardware is shown in Fig. 1. It consists of two components: a pressure sensor array mounted on a belt worn by the mother and an electronic support system.

The final version of the belt evolved over three generations of development, each an improvement over its predecessor. Only the second and third generation belts, shown in Figs. 2 and 3, were used in clinical studies. The second generation belt, containing three sensors in a linear arrangement, was compact enough to be used simultaneously with a commercial ultrasonic monitor. The third generation belt, containing a seven element sensor array to cover the 12-cm diameter range of the fetal heart tone, represents a final practical embodiment.

The electronic support system comprises instrumentation amplifiers to amplify the signals from the sensors; a multiplexer to permit the choice of the best located sensor; a bandpass filter (20-55 Hz); a digital signal processor to implement a linear prediction routine and to yield the FHR in real time; a parallel A/D converter to display the tone on a video display; and a strip chart recorder to indicate the FHR and fetal movement. This system is not yet portabilized and was developed specifically to test the monitor concept. Some of the components serve only to expedite a comparison of the ultrasonic and acoustic NSTs and would not be included in the final version of the portable monitor. Details are given in ref. 1.

Each sensor on the belt is designed to fulfill five functions: signal detection, acceleration cancellation, acoustical isolation, electrical shielding, and electrical isolation of the mother. A cross section and cutaway views of the sensors installed on the second generation belt are shown in Fig. 4. The construction is similar on the third generation belt, except that the electrical conduits are etched on a flexible printed circuit foil.

The internal sensor detects pressure pulses on the maternal abdomen. As recommended by the manufacturer (ref. 2), two PVF2 piezoelectric polymer elements are arranged in a bimorph structure. Here the bimorph operates in the compressional mode with e_{33} the active piezoelectric modulus. Electrical contacts to the Ni electrodes are made with conducting epoxy. The external sensor, identical in construction, is intended to cancel accelerations due to rigid body motion of the mother. A layer of kevlar wool serves to

attenuate signals due to ambient noise. The two sensors are connected differentially to the instrumentation amplifier. The Ni plate attached to the belt is bonded to both bimorphs and assures that both are subjected to the same accelerations. A foil of Cu coated kapton completely surrounds the sensor assembly. When the ground wire connected to the Cu coating contacts earth ground, the ubiquitous 60 Hz interference plummets into the background. A final layer of RTV silicone rubber, completely covering the Cu shield, isolates the mother from earth ground but is acoustically transparent. The belt itself is made of nylon parachute webbing. It does not have to be drawn tightly around the mother, rather requiring only minimal acoustic contact, but must have a sufficiently high modulus to resist displacement by the incident pressure pulses and thus assure adequate compression of the PVF2 foil.

CLINICAL VALIDATION OF THE PORTABLE MONITOR

The fetal non-stress test (NST) is performed routinely in hospitals by means of pulsed Doppler ultrasound. A normal NST requires three separate FHR accelerations of at least 15 beats per minute over its baseline. Each acceleration event is to be stimulated by an associated fetal movement. These three acceleration/movement events are to occur during any 20-minute observation window. When the mother perceives a fetal movement she records this event by pressing a push-button switch. The fetal heart rate measured by the sensor is recorded continuously on a strip chart recorder. The tests described here were conducted on patients who came to the Eastern Virginia School of Medicine, Norfolk, Virginia, for regular appointments, and then volunteered to take a subsequent test with the acoustically based monitor.

Figure 5 shows a FHR recorded simultaneously by an ultrasonic transducer and a sensor on the second generation belt, which were mounted together on the patient. The arrows at the bottom of the strip chart indicate fetal movement (FM). The acceleration in the FHR of about 15 beats per minute following the fetal movements indicate that this is a normal NST. There is good correlation between the ultrasonic and acoustic recordings.

Figure 6 shows a FHR recorded by means of the third generation belt alone. The width of the belt precluded simultaneous mounting with an ultrasonic transducer. A fetal movement is indicated by a spike right on the recording. The FHR acceleration following the FM confirms that the acoustically based monitor is capable of performing the NST reliably.

The acoustically based fetal heart rate monitor offers the advantages of portability, low cost, increased frequency of surveillance, and home-use by the patient with minimal instruction. Finally, the monitor is truly non-invasive since it does not inject energy flux into the developing fetus.

REFERENCES

1. R. A. Pretlow, "Signal Processing Methodologies for an Acoustic Fetal Heart Rate Monitor," M.S. Thesis, Old Dominion University, April, 1991.

2. *Kynar Piezo Film Technical Manual*, Pennwalt Corp., Valley Forge, PA 19482, 1987.

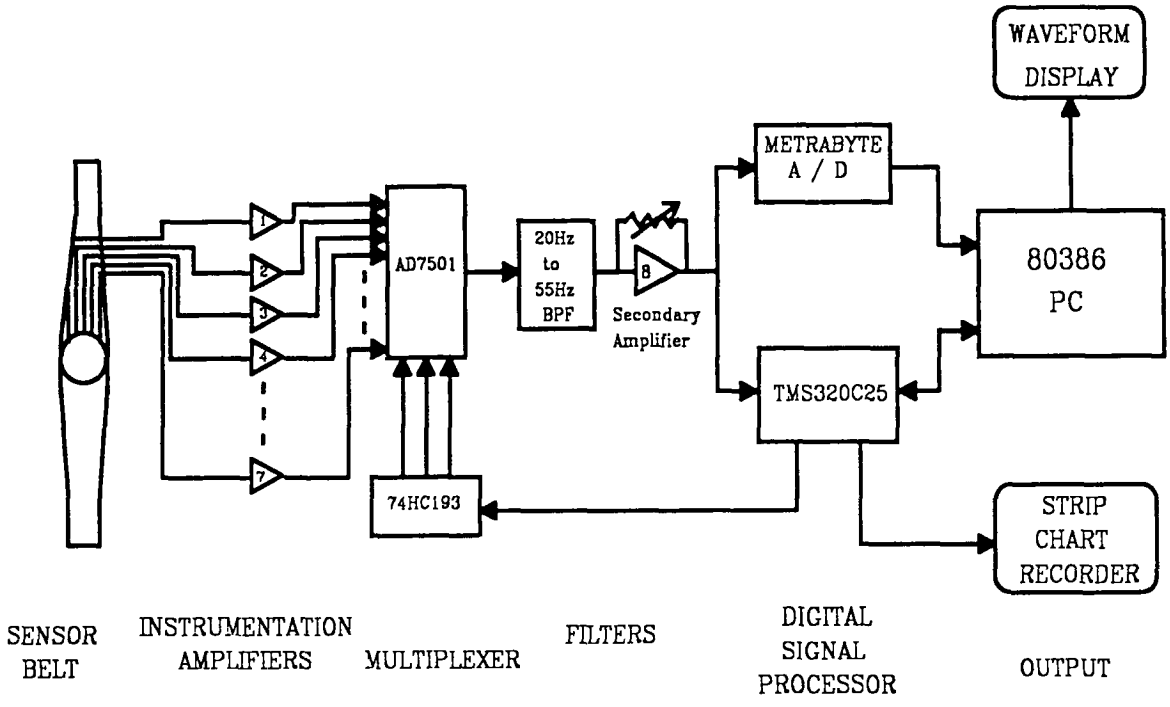


Figure 1. Block diagram of the monitor hardware.

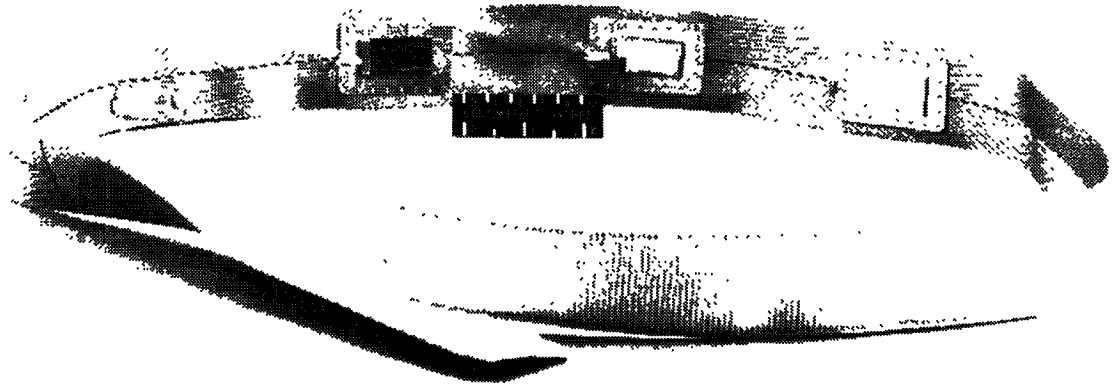


Figure 2. The second generation belt prior to installation of the shielding electrodes.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

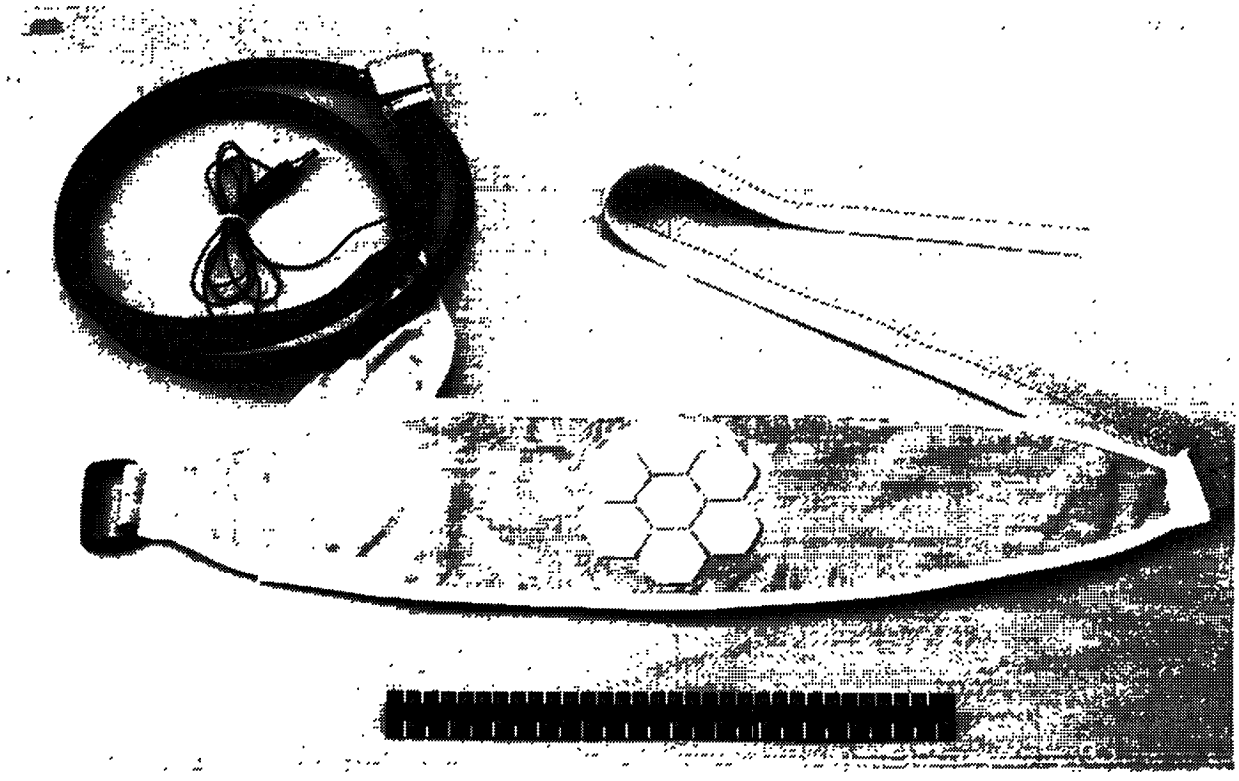


Figure 3. The third generation belt.

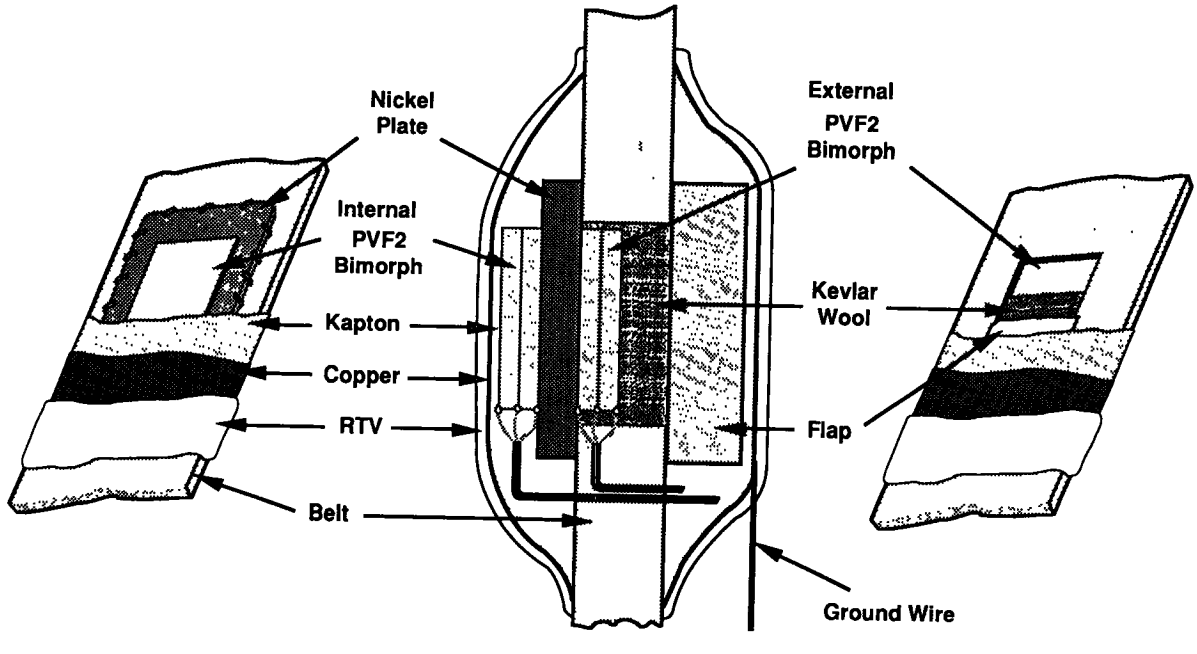
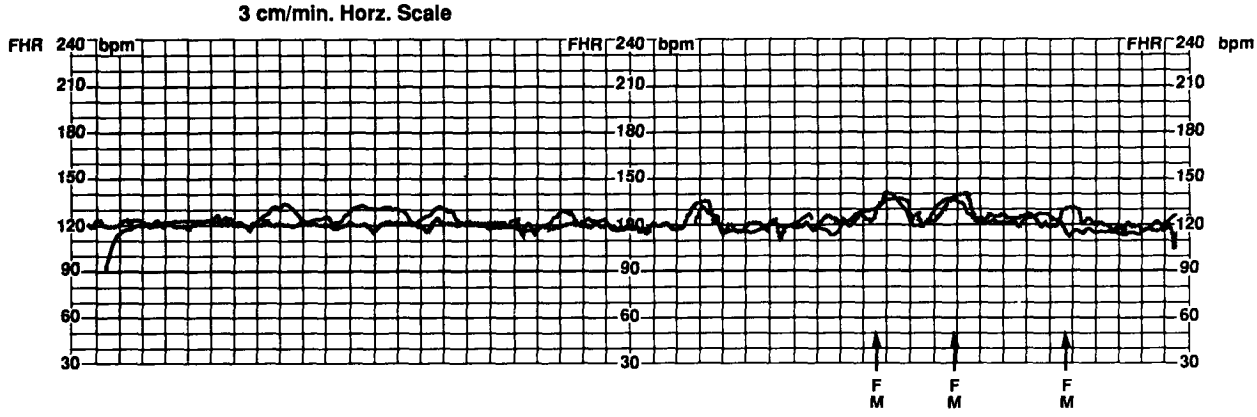


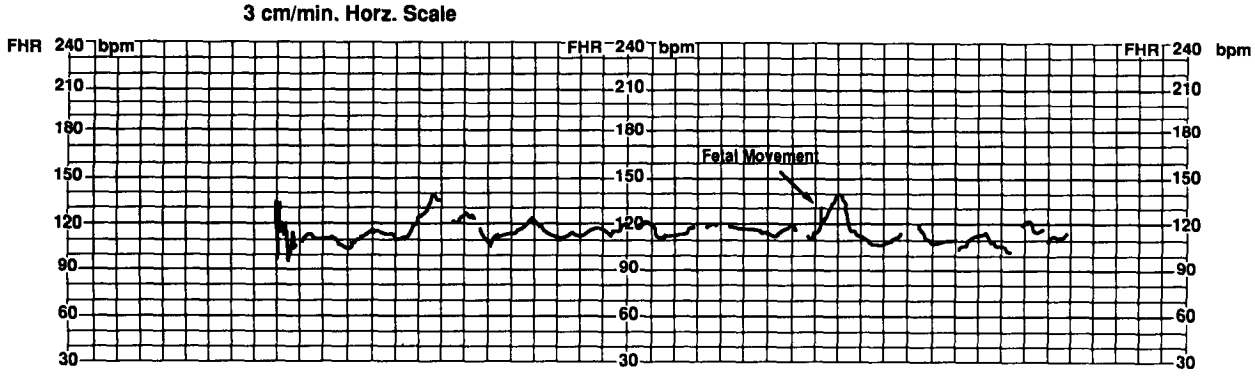
Figure 4. A cross section and cutaway views of the second generation belt.



PATIENT 23

ULTRASONIC ACOUSTICAL ——

Figure 5. A fetal Non-Stress Test recording: comparison between ultrasound and the acoustic monitor.



PATIENT 31

Figure 6. A Non-Stress Test recording acquired from the third generation belt.

N92-22734

SURGICAL FORCE DETECTION PROBE

Ping Tcheng*, Paul Roberts*, Charles Scott* and Richard Prass**

* NASA Langley Research Center
Hampton, Virginia

** Depart of Otolaryngology - Head and Neck Surgery
Eastern Virginia Medical School
Norfolk, Virginia

ABSTRACT

This paper reports the development progress of a precision electro-mechanical instrument which allows the detection and documentation of the forces and moment applied to human tissue during surgery under actual operating room conditions. The pen-shaped prototype probe which measures 1/2 inch in diameter and 7 inches in length was fabricated using an aerodynamic balance. The aerodynamic balance, a standard wind tunnel force and moment sensing transducer, measures the forces and the moments transmitted through the surgeon's hand to the human tissue during surgery. The prototype probe which was fabricated as a development tool was tested successfully. The final version of the surgical force detection probe will be designed based on additional laboratory tests in order to establish the full scale loads. It is expected that the final product will require a simplified aerodynamic balance with two or three force components and one moment component with lighter full scale loads. A signal conditioner has been fabricated to process and display the outputs from the prototype probe. This unit will be interfaced with a PC-based data system to provide automatic data acquisition, data processing and graphics display. The expected overall accuracy of the probe is better than one percent full scale.

INTRODUCTION

The objective of this Technology Utilization Office funded project is to develop a surgical force detection probe suitable for medical applications at the request of Dr. Richard Prass of the Eastern Virginia Medical School in Norfolk, Virginia. Based on Dr. Prass's microsurgical experiences and requirements it was decided that a multi-component strain gaged aerodynamic balance can be modified to meet the objective of the surgical force detection probe. The use of a force detection device provides the ability to monitor the forces applied to human tissue during surgery under actual operating room conditions. Dr. Prass cited the following advantages:

(1) It allows documentation of the usual forces applied during routine surgical procedures. Such documentation has never been reported.

(2) It allows comparison among experienced surgeons and those in training. Such data may provide feedback that may be effectively used during residency training.

(3) When used in conjunction with interoperative neurological monitoring, it will allow correlation of specifically applied forces to monitored nerves that are responsible for nerve injury. These data may lead to new concepts in nerve dissection that improve surgical outcome.

The aerodynamic balance, a standard precision wind tunnel force and moment detection transducer, is used to measure the forces and the moments applied to the test model. Most of the balances used at Langley Research Center (LaRC) are designed to measure three forces and three moments. The prototype surgical force detection probe, shown in Fig. 1, consists of three components: an external housing or a cover shield, an

aerodynamic balance and a clamp adapter for holding interchangeable probe tips. It will measure the forces and moments transmitted through the surgeon's hand to the human tissue during surgery. The prototype unit utilizes an existing six-component aerodynamic balance as a development tool.

MECHANICAL DESIGN

The prototype surgical force detection probe incorporates LaRC's 747 strain-gage balance, a six component aerodynamic balance which measures three forces and three moments. Balance 747 is oversized in loads as it has the following full scale design loads: 30 lbs in normal, 10 lbs in axial, 20 lbs in side, 40 in-lbs in pitch, 10 in-lbs in roll and 20 in-lbs in yaw. It is used only for proof of concepts and for laboratory evaluation to establish the load requirements of the final production version of the surgical force detection probe.

The prototype probe measures 0.5 inch in diameter and 7.0 inches in length, excluding the probe tips. Shown in Fig. 2, the actual probe will have the same diameter as the prototype, however, it will be approximately one inch shorter in length. The prototype probe as well as the final probe consists of the same three main components. An assembly view depicting these three components can be seen in Fig. 3. The clamp adapter is metric and is used to hold various microsurgery probe tips. The cover shield serves as the mechanical ground and is held in the surgeons hand. The transducer contains three strain-gaged measuring beam sections which yield output signals proportional to the loads transmitted from surgeon's hand to the probe tips. This compact probe will allow accurate monitoring of the forces and moments that the surgeon is applying to the patient during microsurgery.

In order to produce a multi-component transducer, such as the surgical force detection probe, many steps must be taken. The design specifications and constraints must be established first before actual mechanical design. Following design, the transducer must be fabricated, strain gaged and calibrated. This series of steps will be discussed in the following.

Design Specifications and Constraints Design specifications and constraints for the surgical force detection probe are listed below:

(1) Design loads: The full scale design loads for the probe have not been firmly established. It is expected that the design loads will be fairly light. One pound in normal and side forces and two inch-pound in rolling moment or torque are the tentative design loads for the actual probe. The exact values will be determined by experimenting with the prototype probe.

(2) Probe size: The size and weight of the probe should be minimized in order not to impede surgeon's operation.

(3) Overload protection: Mechanical protection should be included in the design to prevent accidental overloading of the probe.

(4) Sterilization: The probe must withstand the sterilization process.

(5) Tip insertion method: Interchangeable probe tips must be incorporated in the design.

The first two specifications are met by choosing aluminum as the transducer material and sizing the measuring beams as small as possible. However, the measuring beams must be large enough for strain-gage installation and they must yield sufficient outputs to maintain measurement accuracy. A photograph of the prototype transducer measuring beams is shown in Fig. 4. Two of the measuring beam sections are used to measure normal force, side force, and the rolling moment. The center beam section measures axial force. All of the measuring sections contain parallel beam configurations. Spring constants are calculated for each beam set to determine the system stiffness relative to the magnitude of the measured loads. This procedure is discussed in more detail in reference 1. The beams are designed to have a relatively low spring constant to

measured forces and moments while having higher spring constants to unmeasured forces and moments. This ensures that the transducer will be sensitive to the measured loads and insensitive to the unwanted loads. In order to obtain acceptable sensitivities the measuring beams will be highly stressed during operation. For this reason, the actual probe will be manufactured out of 7075-T6, a lightweight, high strength aluminum which has a 2% yield strength of 72,000 psi [2].

To avoid damaging and to improve the durability of the probe, it was decided that overload protection should be provided. This is difficult to incorporate into the design since the full scale deflection of the measuring beam systems is typically on the order of 0.001 inch. It is difficult to maintain clearance with such a small gap between the metric end of the balance and the cover shield. The method of providing mechanical stops has not yet been determined. To date, the most feasible idea is to provide set screws around the circumference of the cover shield which can be adjusted, prior to probe use, to set the clearance. This process may prove to be impractical to the medical staff. If providing mechanical stops is not feasible, then an electronic alarm will be included into the electronics control box. The alarm will alert the surgeon of potential overload conditions.

To incorporate interchangeable probe tips, a clamp adapter is attached to the front end of the prototype probe using a close diameter fit and a threaded dowel to secure the position. A dowel knocker is used to remove the positioning dowel, allowing disassembly of the adapter from the transducer. The clamp adapter and some examples of probe tips are shown in Fig. 5. However, in most cases, the adapter need not be removed from the transducer to change probe tips. The tapered clamp on the end of the adapter is designed such that probe tips can be quickly interchanged during an operation. To change a probe tip, the adapter is held and the tapered clamp rotated, counterclockwise. By holding the adapter, all of the loads induced during tip replacement are grounded and can not overload the measuring beams. The probe tip is removed and replaced after the tapered clamp is loosened. The clamp is rotated clockwise to tighten against the new tip. When clamped, the surgical force detection probe is ready for use.

The final design requirement on sterilization is currently being investigated. The main question to be answered is the effect of the sterilization process on the service life of the adhesive and gage coating used on the transducer. A strain-gaged test beam will be subjected to prolonged sterilization in ethylene chloride gas. Also, to increase the life of the strain-gages on the actual probe, a thin sheath over the metric to nonmetric joint will be used to shield the gages from debris during an operation. A mandrel has been designed to manufacture a .003 in. walled sheath out of Dow Corning Silastic #Q7-4840 which is a medical grade liquid silicon. As shown in Fig. 2, the sheath will be anchored in the grooves on the forward end of the transducer. The effects of the sheath bridging the metric and nonmetric ends of the probe will be taken into account during calibration.

Fabrication Fabrication of a one-piece multi-component force transducer is a very complicated process. There are many difficult cuts and critical dimensions that must be held to tight tolerances of $\pm$ 0.0005 inch during fabrication. The actual transducer will be manufactured from a single piece of aluminum to eliminate as many joints as possible. One-piece transducers yield smaller zero shifts than multiple piece transducers and are therefore inherently more accurate. A large portion of the machining is performed by Electrical Discharge Machining (EDM) [3]. EDM allows the multiple beam measuring sections to be precisely machined using a single piece of material. Basically, EDM removes metal from the work piece by vaporizing the top surface through an electric spark. A dielectric fluid flushes away the molten metal. This process slowly erodes the metal until the desired dimension is obtained. It would be impossible to fabricate this compact transducer without the EDM process.

Strain-gaging Mounting strain-gages on the measuring beams of the transducer is the next step in production. 5,000 ohm foil strain-gages will be bonded to the measuring beams and interconnected in Wheatstone bridge forms [4]. The 5000 ohm gages will allow larger input voltages to be applied to the bridge, yielding larger signal outputs without heat build up on the measuring beams. The measuring beams are therefore not required to be stressed as highly as beams with lower resistance strain-gages, consequently providing a larger safety factor to the design. The strain-gages are placed such that they are sensitive to the measured component while electrically canceling the unwanted components. This method greatly reduces the interactions between different

components. By carefully selecting the strain-gages and their locations, real time correction of nonlinear interaction errors is unnecessary.

Calibration The final step of production, calibration of the transducer, is one of the most important. Calibrations are most accurate when performed by methods which closely simulate the actual use of the transducer. Calibration hardware is designed to apply dead weight loads similar to those which will be applied to the metric probe tips. Care was taken to make the calibration hardware as light as possible to minimize the initial or tare loads applied to the transducer. The arms can be adjusted for variable load point distances. This allows calibration for various length probe tips. To begin calibration, the probe outputs will be connected to a data acquisition system which consists of a power supply, a scanner, a voltmeter, and a personal computer. The cover shield is positioned such that dead weight loads are applied in line with a force and the moment center. Incremental loads are applied to the calibration arms which are inserted into the tapered clamp. The electrical output signals are recorded and reduced using a short arm calibration method. This is repeated for each force and moment component. First and second order interaction terms are calculated to determine whether they have significant or negligible effects on the accuracy of the probe [5]. If the effects are negligible, then the data can be reduced using only the sensitivities of each component. However, if the effects are significant, then interaction terms must be accounted for, thus making the data reduction more complex. Due to the equal force magnitudes and the calibration procedure, it is expected that the interaction effects will be negligible.

ELECTRONICS DESIGN

Figs. 6 & 7 are two photographs depicting a separate control box which houses the signal conditioning circuits, the front panel controls and the back panel. The control box provides two modes of operation: the manual and CPU modes. This operation will be briefly described in the following.

Modes of Operation

Manual Mode: In this mode, the CPU or a personal computer is not required and the control box will operate as a stand alone unit to process and display the probe outputs. Here, the control box amplifies and filters the probe outputs and displays the signals on bar graphs located on the front panel of the control box. Manual nulling and gain setting will be required for this mode of operation.

CPU Mode In the CPU mode, bridge outputs are routed through a different path in the control box for amplification and filtering. The processed signals are simultaneously transmitted to the CPU and displayed on the same front panel bar graphs. Under software control, the CPU will control offset nulling and gain setting on the programmable amplifier. Additional signal processing and graphic displays by the CPU are planned.

The processed signals are also available for retransmission and recording through the BNC connectors on the back panel for either manual or CPU mode.

The electronics hardware will next be described in the following.

Signal Conditioner The block diagram for a typical channel of the signal conditioner is shown in Fig. 8. It comprises seven major blocks: an isolation amplifier, a null/span adjust, a control switch, a low pass filter, a bar graph display, a multiplying DAC, and a programmable amplifier. As shown, the signal conditioner communicates with a data acquisition system. This interface box remains to be defined. The hardware contained in each block and its function are briefly described in the following using a more detailed schematic diagram shown in Fig. 9.

(1) Isolation Amplifier - An isolation amplifier is commonly used in medical equipment where safe and accurate measurement of a voltage signal is required. The isolation amplifier used employs transformer

coupling and an amplitude modulation technique to provide a complete isolation between input and output signals. The amplification is set at 100. A voltage follower was added behind this isolation amplifier to remove any loading problem from the null/span adjust circuit.

(2) Null/Span Adjust - An operational amplifier is used to provide the null and span capabilities. Null adjust is used to remove output offset which may be caused by tare weight while span adjust allows the operator to increase the output to the full scale value of the output stage. Note that Null/Span Adjust is active only in the manual mode.

(3) Control Switch - A SPDT toggle switch is used to select either the manual or CPU mode. In the manual mode, the output from the Null/Span Adjust is sent to the low-pass filter. In the CPU mode, the output from the programmable amplifier is sent to the low-pass filter.

(4) Low Pass Filter - An active two-pole low-pass Butter-worth filter is used for conditioning the signal coming from the control switch. The corner frequency of the filter is set at 1 Hz.

(5) Bar Graph Display - For visual display two 10-segment LED bar graphs are used to form a 20-segment LED zero center bar graph. Each bar graph is driven by an operational amplifier and by a bar display driver. The driver senses the analog input voltage levels from the low-pass filter and provides a linear analog display on the 10 LEDs. A red LED bar graph is used to display positive output while a green one is for negative output. The bar graph display is on for both manual and CPU modes.

(6) Multiplying DAC - An 8-bit multiplying digital to analog converter is used to generate the offset required to null the programmable amplifier in the CPU mode. It provides an output equal to the product of a fixed reference signal at 1.2 volt and the fractional equivalent of the 8-bit digital word which will be supplied by the CPU. The multiplying DAC block includes two operational amplifiers needed for bipolar operation. The entire block is calibrated to provide a full scale voltage of $\pm$ 2 volts.

(7) Programmable Amplifier - This is a precision instrumentation amplifier with selectable gains of 1, 10, 100 and 500. Two CMOS compatible gain control lines selected by the CPU are used to pick the required gain. The programmable amplifier takes its input from the isolation amplifier, subtracts the output from the multiplying DAC and amplifies the difference by the gain selected.

Front Panel Control As shown in Fig. 6, the front panel of the prototype control box contains the potentiometers for the null/span adjust and the 20-segment LED displays for three probe channels. A ten-turn precision potentiometer is used for span adjustment and a one-turn potentiometer is used for null or zero adjustment. The front panel also contains a power switch, the manual/CPU switch and an auto zero LED display. The LED display is on when the foot switch is depressed.

Foot Switch External to the control box, a foot operated switch is provided to generate a timing pulse. This pulse will signal the CPU to mark the current signal levels and/or null out the programmable amplifiers in the control box. This feature is necessary for the two following reasons: it allows the surgeon to synchronize his surgical operation with external video taping, and it gives the surgeon a record to begin a sequence of steps, when the orientation of the probe is changed. Ideally, this switch should be installed on the cover shield of the force probe for better access by the surgeon.

CONCLUDING REMARKS

In an effort to advance the state of art of surgical instrumentation, the feasibility of using a small aerodynamic balance as a surgical force detection probe which monitors and documents the forces and moments applied to human tissue during surgery under actual operating room conditions has been demonstrated. The pen-shaped prototype probe, approximately 1/2 inch in diameter and 7 inches in length, measures the forces and the moments transmitted through the surgeon's hand to the human tissue during surgery. A prototype probe

using an existing aerodynamic balance as a development tool, as well as the signal conditioner, were fabricated and tested successfully.

The final version of the surgical force detection probe will be designed and fabricated based on laboratory test results using the prototype probe. A fully automated PC-based data system will also be developed for data acquisition and graphics display.

The probe tip can easily be replaced by a pen to convert this device into an instrumented writing tool. Such a tool can be used to monitor the steadiness of handwriting which may be useful to check for soberness in the law enforcement field.

REFERENCES

1. Roberts, P. W., "Development of a Dual Strain Gage Balance System for Measuring Light Loads," 13th International Congress on Instrumentation in Aerospace Simulation Facilities, Gottingen, W. Germany, September 18-21, 1989, pp. 1422-1423.

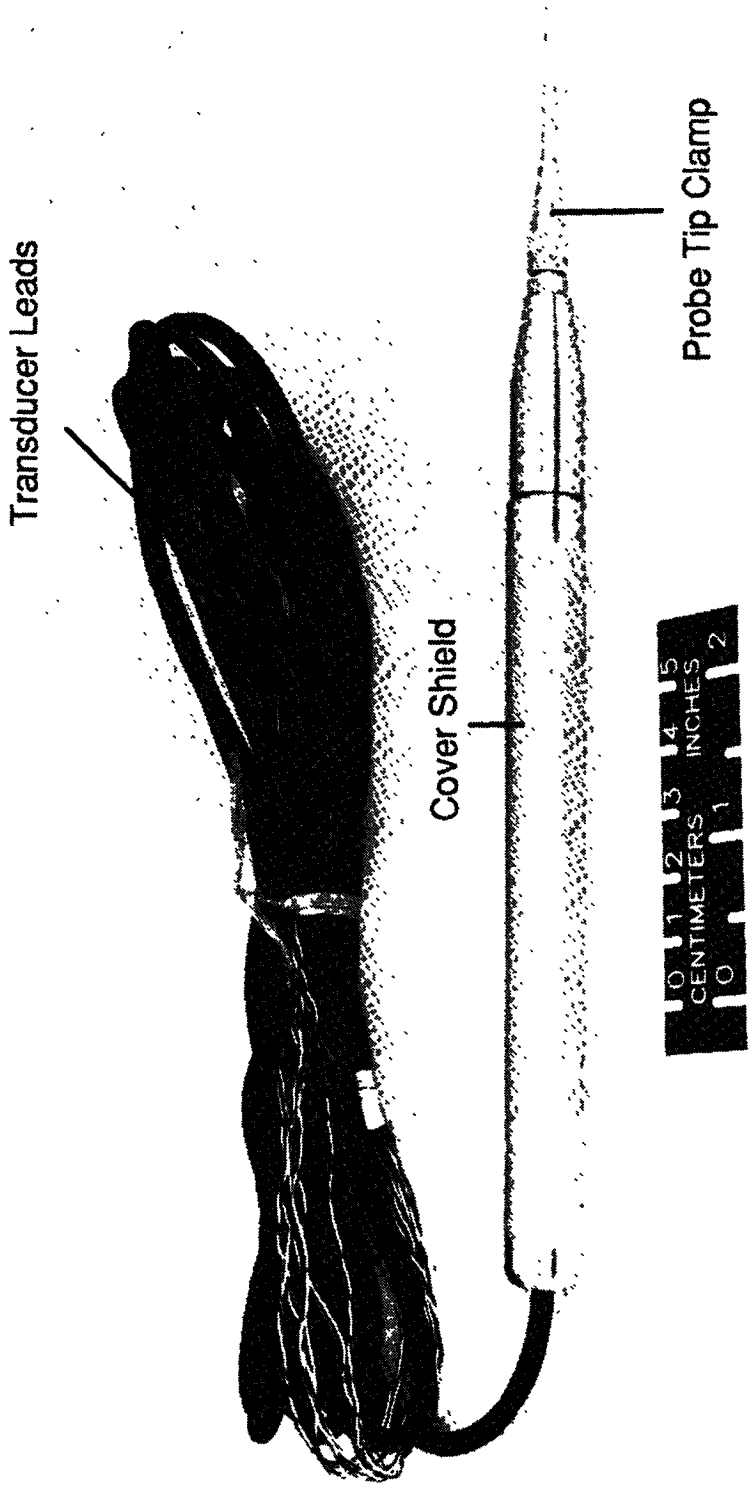
2. Shigley, J. E., and Mitchell, L. D., "Mechanical Engineering Design," 4th ed., McGraw-Hill, New York, New York, 1983, pp. 826.

3. Reda, M. R., "Fundamentals of the EDM Process," EDM Digest, Farmington, MI., March/April 1980, pp. 12-17.

4. Peary, C. C. & Lissner, H. R., "The Strain Gage Primer," McGraw-Hill, New York, New York, 1955, pp. 23-24.

5. Hansen, R. H., "Evaluation and Calibration of Wire-Strain-Gage Wind Tunnel Balances under Load," Presented at Wind Tunnel And Model Test Panel Of Advisory Group For Aeronautical Research And Development, Rome, Italy, February 20-25, 1956, pp. 4-8.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



L-91-14037

Fig. 1. The Prototype Probe

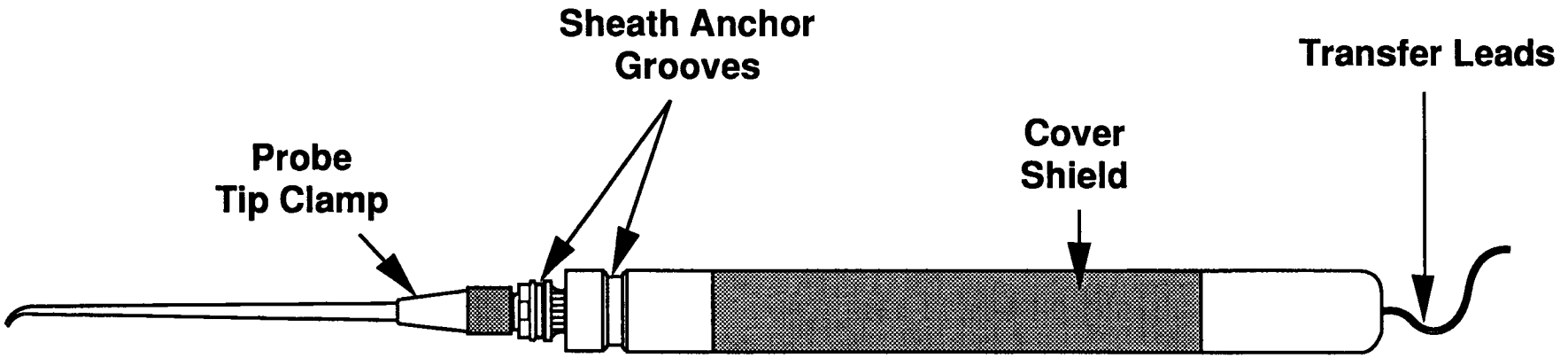


Fig. 2. An Assembley View of the Production Probe

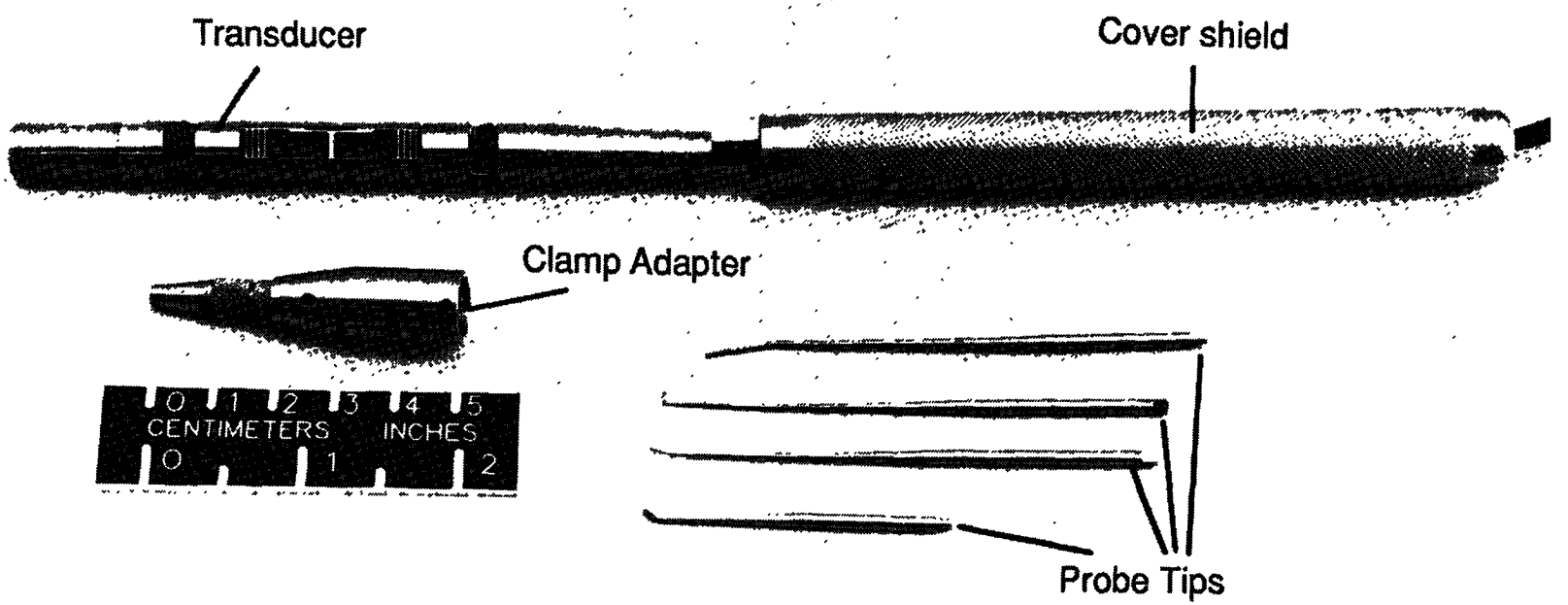


Fig. 3. The Disassembled Prototype Probe

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Fig. 4. Measuring Beam Sections of Balance 747

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

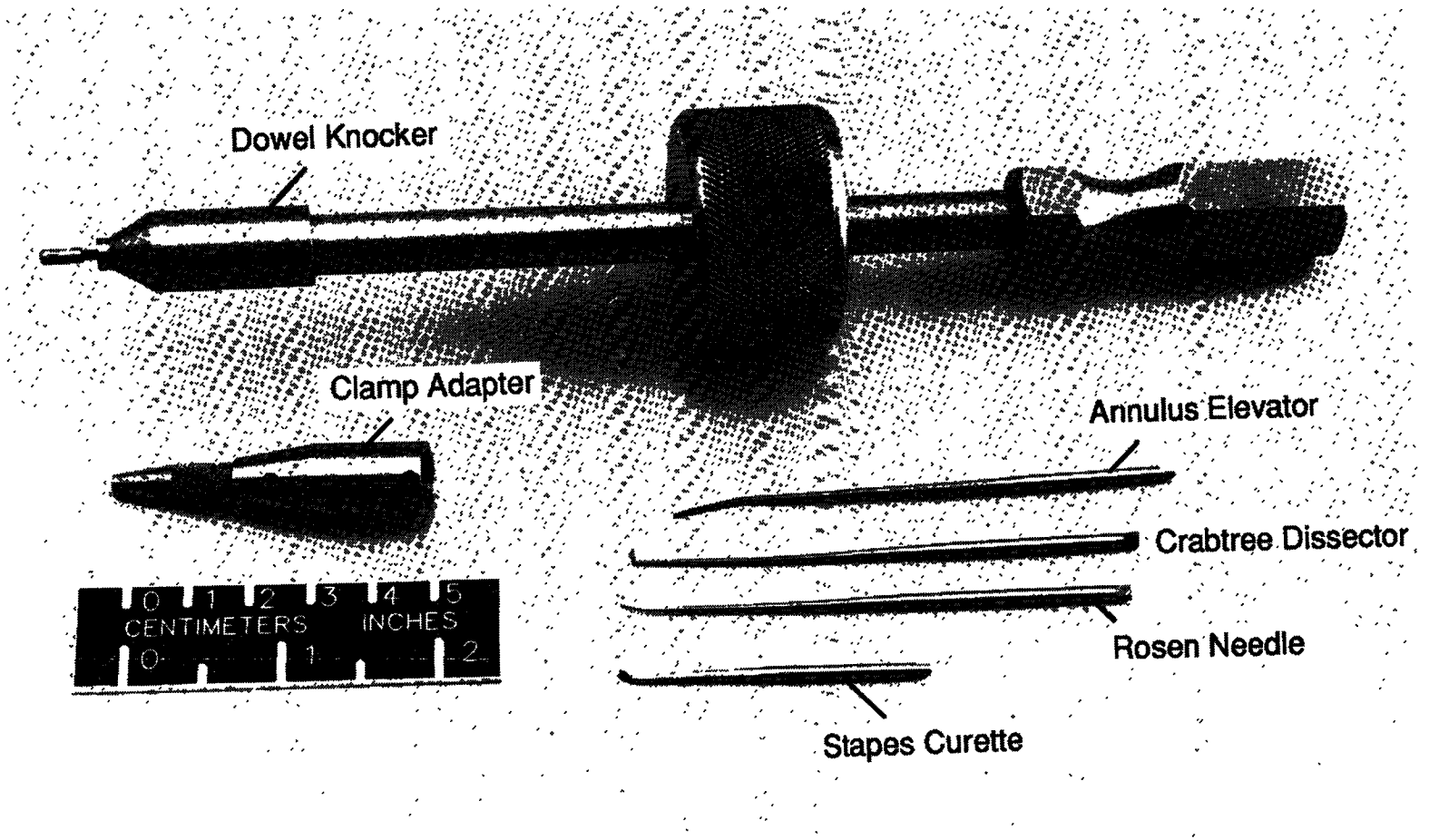


Fig. 5. The Clamp Adapter, dowel Knocker,and Probe Tips

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

L-91-1436

Fig. 6. The Control Box Front Panel

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Fig. 7. The Back Panel and Internal Circuitry of the Control Box

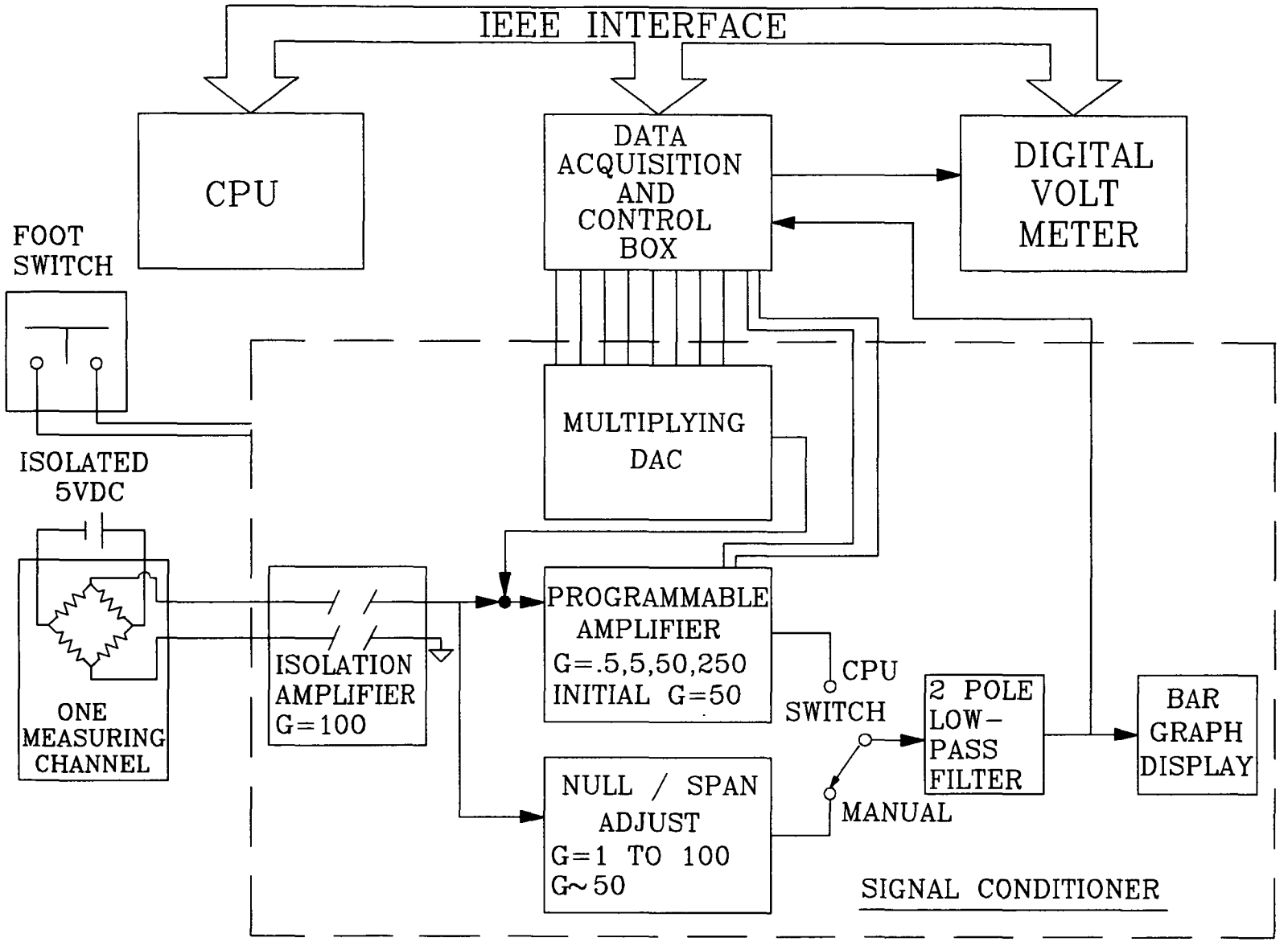


Fig. 8. The Signal Conditioner Block Diagram

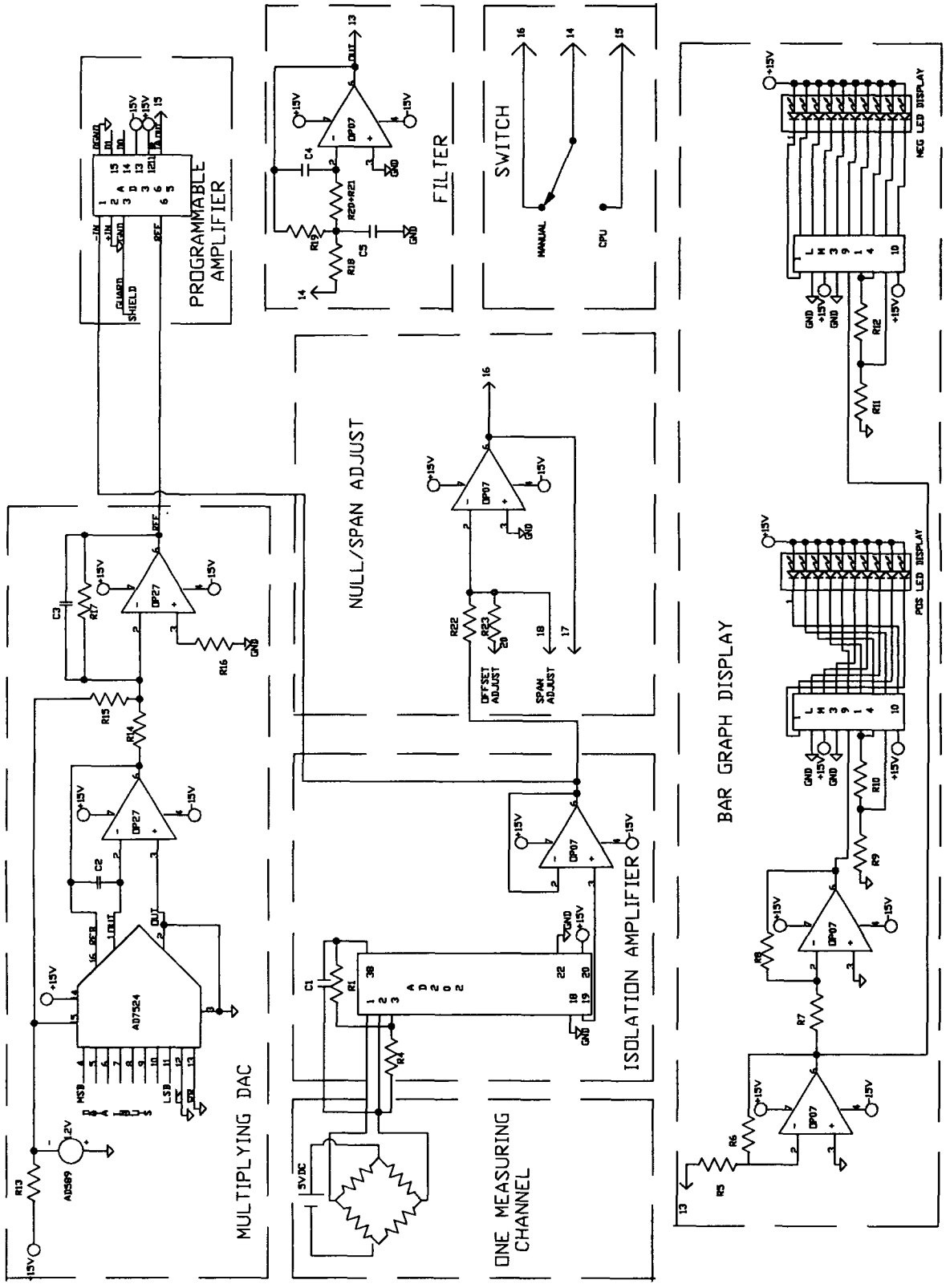


Fig. 9. A Schematic Diagram for the Signal Conditioner

N92-22735

DYNAMIC INTER-LIMB RESISTANCE EXERCISE DEVICE FOR LONG-DURATION SPACE FLIGHT

Douglas F. Schwandt, Donald E. Watenpaugh, Scott E. Parazynski, and Alan R. Hargens

Life Science Division (239-11)
NASA Ames Research Center
Moffett Field, CA 94035-1000

ABSTRACT

Essential for fitness on Earth, resistive exercise is even more important for astronauts, who must maintain muscle and bone strength in the absence of gravity. To meet this need, designers and scientists at NASA Ames Research Center, Life Science Division, have worked to develop more effective exercise devices for long-duration exposure to microgravity [1]. One of these concepts is the Inter-Limb Resistance Device which allows the subject to exercise one limb directly against another, strengthening muscle groups in the arms, legs and back. It features a modular harness with an inelastic cable and instrumented pulley. Forces similar to other high resistance exercise equipment are generated. Sensors in the pulley measure force and velocity for performance feedback display and data acquisition. This free-floating apparatus avoids vibration of sensitive experiments on board spacecraft. Compact with low mass, this hardware is also well suited for a 'safe haven' from radiation on board Space Station Freedom, and may prove useful in confined environments on Earth, such as Antarctic stations, submarines, and other underwater habitats. Potential spin-offs of this technology include products for personal strengthening and cardiovascular conditioning, rehabilitation of hospital patients, fitness exercise for the disabled, and retraining after sports injuries.

INTRODUCTION

This paper describes the concept, features, and preliminary trials of a simple, low mass resistive exercise device. The concept of working one limb against another limb is common to many activities in our daily lives. For instance, when we put on shoes, we pull with our hands while pushing with our foot. In the early part of this century, Charles Atlas first systematized this activity as a simple way to exercise a variety of muscle groups without any equipment, and this technique was popularly termed "dynamic tensioning". For example, in one dynamic tensioning exercise, the subject presses one hand against the other hand with arms parallel to the ground, elbows pointing laterally, while moving the hands and arms as a unit from side to side. More recently, inventors have patented mechanisms to facilitate inter-limb resistance. One mechanism transfers forces and motion through levers and chains [2], and another through racks and gearing [3]. The Inter-Limb Resistance device builds upon these mechanisms providing a simple, light-weight system to help transfer forces from one part of the body to another part, and to measure and display the forces and limb velocities.

DESCRIPTION OF DEVICE

Modular Harness, Rope and Instrumented Pulley

Instead of the heavy, rigid steel frames associated with traditional weight stack machines found in fitness clubs, the Inter-Limb Resistance device weighs less than 10 pounds (not including the computer and display) and uses a soft conforming seat harness with shoulder straps, shoulder straps, ergonomic handles, and footplates to interface with a flexible cable and instrumented pulley. The footplates are secured to the feet with quick release straps. As the subject exercises, energy is delivered directly through the inelastic cable and pulley(s) from one limb to another. For instance, during biceps curls, each arm generates an almost identical upward force against each end of the cable, while one arm moves down and the other arm moves up (Fig. 1a). The instrumented pulley is connected to the footplates; thus, the legs resist the forces from both arms. The arm velocity, the combined force from the two arms, and the instantaneous power (force x velocity) are all displayed to the subject as motivational feedback. Data is also digitally stored for further analysis.

At all times during the exercise, the subject has full control of the speed and direction of limb motion and magnitude of muscle force. For instance, the subject may choose to exercise with low force at the extremes limb position to increase flexibility, while generating very high force at the high leverage portion of the stroke to increase muscle strength. Forces similar to other high resistance exercise equipment may be generated as the subject is only limited by his or her volition. The subject's direct control of the forces and velocity, combined with the lack of stored energy, provide for a safe exercise that can be stopped at anytime if the subject feels pain from over exertion.

Concentric, Eccentric and Isometric Exercise

In everyday activities, our muscles contract concentrically, eccentrically and isometrically, depending upon whether the muscle fibers are shortening, lengthening or not changing length, respectively, during muscle activation. For example, when we jump, our leg muscles contract concentrically. When we land, the same muscles need to contract eccentrically to decelerate the body. While stooped to work on something near the floor, those muscles are contracting isometrically to hold the position. During space flight, astronauts experience very little eccentric muscle contraction, especially with their legs, as they do not need to resist gravity. Therefore, it is important to exercise muscles eccentrically during extended space travel [4, 5] to prepare astronauts for return to Earth, or for landing on Mars, when they will need to resist gravity again.

Concentric, eccentric and isometric muscle exercise is possible with the Inter-Limb Resistance device. For example, during biceps curls, the biceps in the arm moving up is contracting concentrically (muscle fiber shortening during activation) as it progresses against the downward force of the cable. The biceps in the arm moving down is contracting eccentrically (muscle fiber lengthening during activation) as it is pulled down by the same downward force of the cable. Essentially, the biceps in the arm moving up is performing work on the biceps in the arm moving down. Contractions with no motion are isometric and may be performed at any position in the range of motion.

Comparison to Other Types of Resistance Exercise Devices

By transferring forces from one limb to another, there is no need for energy storage such as lifting a weight stack to store potential energy or compressing a spring to store strain energy. Such storage of energy on other devices makes it possible to have an eccentric phase on the return stroke, such as letting the weight stack down after a leg press. However, the eccentric phase is always present in one of the limbs with the Inter-Limb Resistance device, when the contralateral muscles are contracting concentrically. Other devices use a friction drum to absorb energy from concentric muscle contractions. Such energy is unavailable for eccentric muscle contractions and is essentially wasted. In addition, some types of exercise equipment require external power, such as an active treadmill. The Inter-Limb Resistance device requires only a small amount of power for the instrumentation.

Modular Components Accommodate Multiple Exercises

With modular components it is possible to reconfigure the system easily to accommodate a variety of arm and leg exercises. For instance, in one configuration, the subject performs a reciprocating leg press (Fig. 1b). Simple adjustment of cable length accommodates various individual sizes, or permits adjustment for varying initial joint angles. For example, by releasing cable, both legs have a more extended initial position. With legs nearly equally extended, hip and knee joints can be held in one position while the foot dorsiflexes and plantarflexes, isolating the soleus and gastrocnemius muscle groups. This is particularly important to astronauts as they often experience atrophy in these muscle groups.

A quick conversion of the harness configuration and repositioning of the instrumented pulley allows arm exercises. A force spreader bar can be attached to the footplate pulley blocks connecting the feet and forming a bridge from above each forefoot. The instrumented pulley is then attached to the spreader bar as a base for arm exercises. To change to another arm exercise, such as military press, the subject quickly adjusts the initial cable length by repositioning the line gripping handles commonly used in sailboat rigging.

Because the arm exercises are resisted at the feet, the forces and moments generated during the several arm exercises must be supported by the back, strengthening important back muscle groups as well. Another conversion facilitates leg abduction (moving the legs apart against force), where one end of the cable is held in the subject's hand resisting the abduction.

Force and Motion Sensors in the Instrumented Pulley

Sensors in the instrumented pulley (Fig. 2) measure force and motion for performance feedback display and data acquisition. Force is measured with a 750 pound load cell (Sealed Super-Mini Load Cell, Interface Inc., Scottsdale, AZ). The velocity of the flexible cable is continually determined with an incremental optical encoder (Model 815, Litton, Chatsworth, CA). The encoder detects the angular motion of the pulley shiv, thereby proportionally measuring the linear motion of the cable as it passes through the pulley. An IBM 386 computer acquires and displays the pulley force and cable velocity data in conjunction with the data from an accelerometer. The computer interface was designed using a commercial software development package for data acquisition, control and display (LabWindows, National Instruments, Austin TX). The display is presented on an IBM VGA color monitor and includes graphs, numbers and bar meters of the instantaneous force, velocity and power (the product of force times velocity) for each bout of exercise. The G-level from the accelerometer, useful during zero-G testing on board NASA's KC-135, is presented as a digital number from zero to two Earth gravitational equivalents.

The motion and forces sensed at the instrumented pulley can be related to limb motion through the geometry of the flexible cable and harness for each exercise. For instance, during reciprocating leg presses, a single cable is connected at one side of the seat harness, runs down the outside of the leg, through the foot pulleys, back up between the legs, through the instrumented pulley which is connected to the lower back of the seat harness, and then symmetrically interfaces with the other leg. During exercise, the force developed by each leg is resisted approximately equally by the cable on the outside of the leg which attaches to the seat harness, and by the cable on the inside of the leg which runs through the instrumented pulley. Therefore, the instrumented pulley senses only half of the force developed by each leg. A swivel allows the instrumented pulley to align with changes in force direction.

During arm exercise, the instrumented pulley resists the total force from each arm, because the cable ends at each handle. Therefore, it is currently necessary to distinguish exercises when reading the measured load cell forces to know how the force data relate to the contribution of a particular limb during exercise. We plan to incorporate configuration information into the human-computer interface, so the subject can simply indicate the name of the exercise and then receive meaningful feedback during the exertion.

In tension, the cable is the most efficient way to transfer forces, as the stresses are distributed evenly across the cable cross-section. In other words, very little mass is required to do the same function that may be performed alternatively by a linkage or geared drive train. Another advantage of the cable is its flexibility, making it convenient to fold and store with the flexible harness as a lightweight, compact multi-exercise device. The modularity encourages experimentation with other configurations, and the design is expected to continue to evolve with use.

METHODS

To evaluate its performance in actual, short-duration microgravity, a first fully-functional prototype of the Inter-Limb Resistance device was tested on board NASA's KC-135 parabolic flight in March, 1991. To achieve near zero-G free fall, the KC-135 climbs to approximately 38,000 feet elevation and then dives to approximately 22,000 feet creating a parabolic trajectory that repeats typically forty times (4 sets of 10). Over the top of each parabola, subjects are essentially falling in a parabolic curve experiencing a period of approximately 25 seconds of actual microgravity. As the plane pulls out at the bottom of the dive, approximately 1.8-G is felt before the next zero-G period.

With four subjects each exercising during one of the four sets of 10 parabolas, a variety of arm and leg exercises were performed during the simulated microgravity on board the KC-135 flight. They included leg press, running in place, calf press (ankle extension), leg abduction, biceps curl, arm rowing, and military press. Arm exercises and leg exercises resisted through the shoulder harness recruited back muscle groups for spinal loading, as well. All exercises were performed at sub-maximal effort.

RESULTS

The Inter-Limb Resistance exercise device performed very well during its first "shake-down" test in the short duration microgravity on board the KC-135. Three subjects generated peak forces ranging from 640 to 1156 N (144 to 260 lbs) during the leg press (sum of force from both legs). Two subjects performed the calf press and generated

peak forces of 702 to 782 N (158 to 176 lbs). The biceps curl exercise produced up to 623 N (140 lbs) for one subject. Forces approaching peak values were sustained throughout the range of motion for all exercises. Subjects successfully donned, reconfigured, and removed the Inter-Limb Resistance device during microgravity tests.

DISCUSSION

Ground-based studies are planned to validate the efficacy of the Inter-Limb Resistance device as a countermeasure against muscle atrophy due to long-term inactivity or exposure to microgravity. Such studies will investigate basic muscle physiology. For example, the device may be used to isolate muscle groups in the lower leg, employing measurements of intramuscular pressure and electromyography to compare the training effects of eccentric and concentric muscle contractions.

Future generations of this device will be more self-contained, incorporating a miniature full screen display (Private Eye, Reflection Technology, Waltham, MA) for performance feedback, replacing the current dependence on full-sized CRT displays. Ultimately, the system hardware will feature its own dedicated microprocessor as well, requiring no umbilical connection to a separate computer system during exercise. With such a microprocessor, special features may be programmed. For example, protocols may be presented to the subject through a compact display permitting isokinetic, isotonic, and isometric contractions for a variety of arm, leg and back exercises. To accomplish this, the subject will follow a prescribed target value for force and/or velocity throughout the range of motion.

Potential markets for the Inter-Limb Resistance device include those for personal training at home and for sport fitness centers. Also, its application in rehabilitation is being explored in collaboration with the Palo Alto VA Rehabilitation R&D Center.

ACKNOWLEDGEMENTS

This research and development effort has been supported by the Director's Discretionary Fund at NASA-Ames Research Center. KC-135 flight tests were supported by the Exercise Countermeasures Program at NASA-Johnson Space Center, under the direction of Michael Greenisen, Ph.D., in cooperation with KRUG International and the Reduced Gravity (Zero-g) Program, Ellington AFB. Prototyping facilities and program support for this collaborative effort were provided by the Department of Veteran Affairs' Rehabilitation R&D Center in Palo Alto, California.

Conceptual and technical support for this project were provided by John Kiowski (Software Engineer, Krug International), Russ Hays (Electronics Engineer, Krug International), Robert Whalen, Ph.D. (Research Scientist, NASA-ARC), James Anderson (Modelmaker/Machinist, Palo Alto VA Rehabilitation R&D Center), Uwe Meyer, Ph.D. (National Research Council, NASA-ARC), Lin Liang (NASA-ARC contractor and Stanford University Ph.D. candidate). Consultation in preparation for KC-135 trials was provided by Gita Murthy (Physiologist, Bionetics at NASA-ARC). Data from KC-135 trials were analyzed by Karen Hutchinson (Laboratory Assistant, Bionetics at NASA-ARC), and Richard Ballard (Laboratory Technician, San Jose State University Cooperative Program with NASA-ARC).

REFERENCES

1. SCHWANDT, D.F., R.T. WHALEN, D.E. WATENPAUGH, S.E. PARAZYNSKI, AND A.R. HARGENS. Development of exercise devices to minimize musculoskeletal and cardiovascular deconditioning in microgravity. *The Physiologist*, Vol. 34, No. 1, Suppl., pp. S189-S190, 1991.
2. JONES, A.A. Exercising apparatus and method. US Patent #4,493,485, 1985.
3. PITKANEN, A.R. Computer directed exercising apparatus. US Patent #4,556,216, 1985.
4. DUDLEY, G.A., P.A. TESCH, B.J. MILLER, AND P. BUCHANAN. Importance of eccentric actions in performance adaptations to resistance training. Aviat. Space Environ. Med., 62:543-50, 1991.
5. HARGENS, A.R., S. PARAZYNSKI, M. ARATOW, AND J. FRIDEN. Muscle changes with eccentric exercise: implications on Earth and in space. *Advances in Myochemistry*, 2:299-312, 1989.

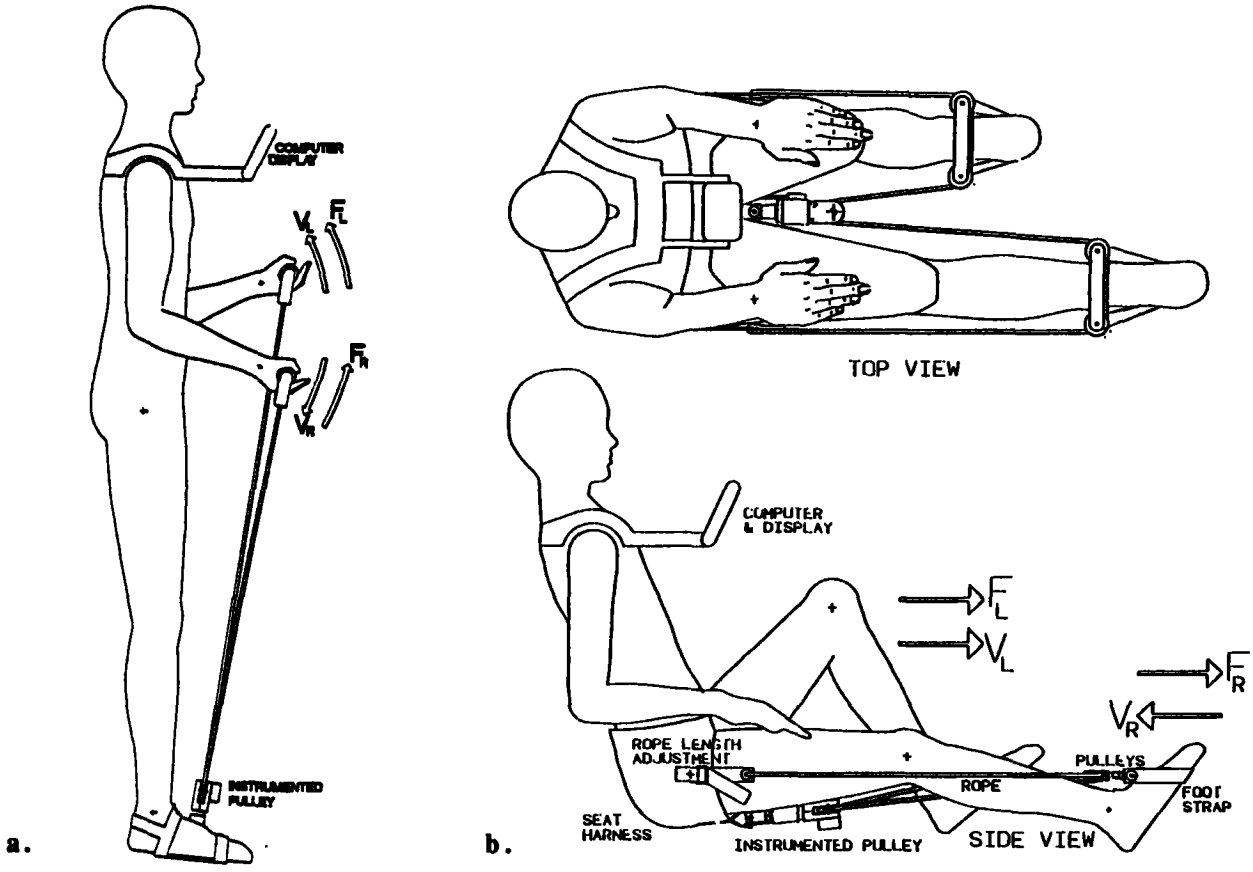


Figure 1. a) During Inter-Limb Resistance to strengthen biceps, the force generated by one arm is resisted by the other. The sum of the force from each arm is resisted at the feet with an instrumented pulley. b) During leg press exercise, the force (F_L) generated during lower body exercise by the subject's left leg as it extends concentrically (velocity, V_L) is delivered through a cable and instrumented pulley to the right leg, which develops an equal resistive force (F_R) during eccentric contraction (V_R). Although not shown, a shoulder harness is used to help resist leg forces and load the spine.

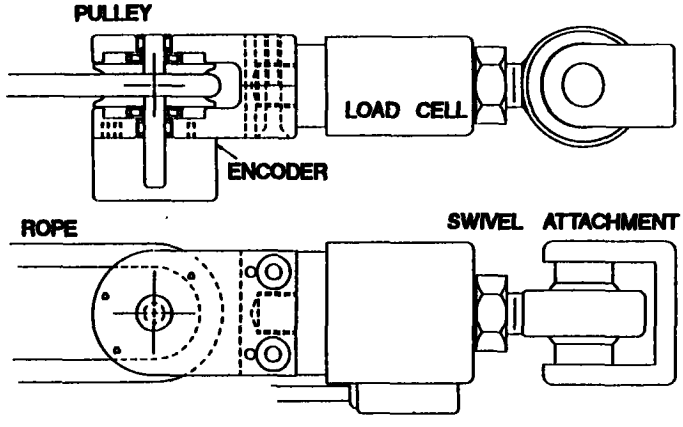


Figure 2. Schematic drawing of the instrumented pulley showing the load cell and the optical encoder. For all exercises, the magnitude and direction of arm or leg velocities and forces are controlled by the subject through the range of motion. Performance data are displayed to the subject.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1991	Conference Publication

4. TITLE AND SUBTITLE

Technology 2001
Volume 2

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Technology Utilization Division

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA CP-3136

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited

Subject Category 99

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Proceedings from the Technical Sessions of the Technology 2001 Conference and Exposition, December 3-5, 1991, San Jose, CA. Volume 2 features 60 papers presented during 30 concurrent sessions.

14. SUBJECT TERMS

technology transfer
computer technology
materials science
robotics

15. NUMBER OF PAGES

537

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

National Aeronautics and
Space Administration
Code NTT

Washington, D.C.
20546-0001

Official Business
Penalty for Private Use, $300

SPECIAL FOURTH-CLASS RATE
POSTAGE & FEES PAID
NASA
Permit No. G-27

POSTMASTER: If Undeliverable (Section 158 Postal Manual) Do Not Return